

DTIC FILE COPY

2

AGARD-AG-182

AGARD-AG-182

AD-A184 834

AGARDograph No. 182

The Practical Assessment of Pilot Workload

DISTRIBUTION STATEMENT A
Approved for public release
Distribution unlimited

DTIC
ELECTE

JUL 1987

D

DISTRIBUTION AND AVAILABILITY

COMPONENT PART NOTICE

THIS PAPER IS A COMPONENT PART OF THE FOLLOWING COMPILATION REPORT:

TITLE: The Practical Assessment of Pilot Workload: Flight Mechanics Panel
of AGARD.

TO ORDER THE COMPLETE COMPILATION REPORT, USE AD-A184 834.

THE COMPONENT PART IS PROVIDED HERE TO ALLOW USERS ACCESS TO INDIVIDUALLY AUTHORED SECTIONS OF PROCEEDING, ANNALS, SYMPOSIA, ETC. HOWEVER, THE COMPONENT SHOULD BE CONSIDERED WITHIN THE CONTEXT OF THE OVERALL COMPILATION REPORT AND NOT AS A STAND-ALONE TECHNICAL REPORT.

THE FOLLOWING COMPONENT PART NUMBERS COMPRISE THE COMPILATION REPORT:

AD#: P005 629 thru P005 643 AD#: _____
AD#: _____ AD#: _____
AD#: _____ AD#: _____

Accession For	
NTIS GRA&I	<input checked="" type="checkbox"/>
DTIC TAB	<input checked="" type="checkbox"/>
Unannounced	<input type="checkbox"/>
Justification	
By _____	
Distribution/	
Availability Codes	
Dist	Avail and/or Special
A-1	

**DTIC
ELECTE
OCT 09 1987
E**

DTIC FORM 463
MAR 85

This document has been approved
for public release and under the
distribution is unlimited.

OP1: DTIC-TID

AGARD-AG-282

NORTH ATLANTIC TREATY ORGANIZATION
ADVISORY GROUP FOR AEROSPACE RESEARCH AND DEVELOPMENT
(ORGANISATION DU TRAITE DE L'ATLANTIQUE NORD)

AGARDograph No. 282

THE PRACTICAL ASSESSMENT OF PILOT WORKLOAD

Edited by

Alan H. Roscoe, MD
Britannia Airways Limited
Luton Airport
Bedfordshire LU2 9ND
United Kingdom



Accession For	
NTIS CRA&I	<input checked="checked" type="checkbox"/>
DTIC TAB	<input type="checkbox"/>
Unannounced	<input type="checkbox"/>
Justification	
By	
Distribution /	
Availability Codes	
Dist	Avail and/or Special
A-1	

This AGARDograph has been prepared at the request of the Flight Mechanics Panel of AGARD.

THE MISSION OF AGARD

The mission of AGARD is to bring together the leading personalities of the NATO nations in the fields of science and technology relating to aerospace for the following purposes:

- Exchanging of scientific and technical information;
- Continuously stimulating advances in the aerospace sciences relevant to strengthening the common defence posture;
- Improving the co-operation among member nations in aerospace research and development;
- Providing scientific and technical advice and assistance to the Military Committee in the field of aerospace research and development (with particular regard to its military application);
- Rendering scientific and technical assistance, as requested, to other NATO bodies and to member nations in connection with research and development problems in the aerospace field;
- Providing assistance to member nations for the purpose of increasing their scientific and technical potential;
- Recommending effective ways for the member nations to use their research and development capabilities for the common benefit of the NATO community.

The highest authority within AGARD is the National Delegates Board consisting of officially appointed senior representatives from each member nation. The mission of AGARD is carried out through the Panels which are composed of experts appointed by the National Delegates, the Consultant and Exchange Programme and the Aerospace Applications Studies Programme. The results of AGARD work are reported to the member nations and the NATO Authorities through the AGARD series of publications of which this is one.

Participation in AGARD activities is by invitation only and is normally limited to citizens of the NATO nations.

The content of this publication has been reproduced directly from material supplied by AGARD or the authors.

Published June 1987

Copyright © AGARD 1987
All Rights Reserved

ISBN 92-835-1546-3



Printed by Specialised Printing Services Limited
40 Chigwell Lane, Loughton, Essex IG10 3TZ

PREFACE

In 1980 a Pilot Paper proposing a "Further Study of Pilot Workload" was submitted to the Flight Mechanics Panel of AGARD. The paper pointed out that despite conferences, working groups, and AGARDographs devoted to the subject little progress has been made towards formulating a readily acceptable definition of the term pilot workload nor towards recognising suitable techniques for assessing levels of workload.

It was concluded that:

- i) Pilot workload is recognised as an important parameter.
- ii) It is evidently difficult to define and use the parameter in a way that is acceptable to all who agree on its importance.
- iii) A further effort is required to try to improve the situation.

The following paragraphs describing the Scope of the Study have been extracted from the original proposal.

"It is believed that in most cases the research worker who is required to make measurements of pilot workload does so with the equipment and facilities that are available in his own laboratory or establishment — indeed sometimes these are equipment that he has CREATED in his own laboratory. The net result is as many techniques, interpretations and definitions as there are research workers — each of whom is usually addressing his own piloting task or sub-task in a way which exercises and makes best advantage of his own methods etc. The result, as graphically exemplified by AGARDographs 233 and 246, is a universal inability to draw any comparisons between the work and conclusions of the many investigators in the field because not only are different methods being used but also different tasks are being addressed."

"By using carefully chosen and precisely defined piloting tasks it is intended that this study will provide the means of collecting, collating and comparing the methods, techniques, interpretations, opinions, and even definitions of specialists who have experience in the field of pilot workload."

"Participants, therefore, are asked to provide a detailed account of the methodology they would employ in assessing workload levels for one or more of these tasks (see Appendices). It is hoped that participants will also identify the limitations of their technique."

The proposal was accepted by the Flight Mechanics Panel and in May 1981 individuals and organisations known to be interested in pilot workload were invited to participate in the study.

The concept of a small and deliberately bounded study was particularly well received and fifteen people expressed their firm intention to participate. However, despite the initial enthusiasm only five contributions had been received by the end of 1982 and it became clear that for various reasons the original idea of a 'paper study' was not going to be fulfilled within the time scale. It was accordingly recommended to the Flight Mechanics Panel that the study in its original form be abandoned. Because of the substantial interest shown in the study the Panel decided in 1984 that an attempt should be made to produce an AGARDograph containing as much as possible of the information hoped for in the study. Although many of the chapters do not conform entirely to the original idea of a "deliberately bounded study" this AGARDograph should provide a useful guide for the person wishing to assess pilot workload for practical reasons; it is not written for the research scientist interested solely in laboratory experiments.

The first chapter consists of a brief introduction to the subject of pilot workload together with an overview of current techniques for assessing levels of workload with particular reference being made to those of practical importance. Each of the remaining fifteen chapters describes one or more techniques presently available — or likely to become available in time. Several of these techniques have been used in practice with some success; other techniques, with varying degrees of development show promise for the future and are therefore also of interest. Hopefully, the reader will be able to find a technique or, more likely a combination of techniques, to suit his or her purpose.

The Flight Mechanics Panel is very grateful to all the many contributors who have given generously of their time to produce this valuable contribution to this important topic. Particular credit goes to the Editor, Dr A H Roscoe, who has used his expertise in this field together with his practical experience of flying to coordinate individual papers into a unique guide to this subject

A.A. WOODFIELD
J.F. RENAUDIE
Members, Flight Mechanics Panel

CONTENTS

	Page
PREFACE	iii
Chapter 1 Introduction by A.H. Roscoe	1
Chapter 2 In-Flight Workload Assessment Using Embedded Secondary Radio Communications Tasks by C.A. Shingledecker	11
Chapter 3 Use of Timeline Analysis to Assess Crew Workload by G. Stone, R.K. Gulick and R.F. Gabriel	15
Chapter 4 Pilot Subjective Evaluation of Workload During a Flight Test Certification Programme By F. Ruggiero and D.M. Fadden	32
Chapter 5 The use of Subjective Workload Assessment Technique in a Complex Flight Task by F.V. Schick and R.L. Hann	37
Chapter 6 Workload Methodology by E. Dorchin and C.D. Wickens	42
Chapter 7 Mental Workload Measurement in Operational Aircraft Systems: Two Promising Approaches by M. Biferno	44
Chapter 8 Cortical Evoked Response and Eyeblink Measures in the Workload Evaluation of Alternative Landing System Displays by R.D. O'Donnell and G. Wilson	52
Chapter 9 In-Flight Assessment of Workload Using Instrument Scan by J.R. Tole and R.L. Harris, Sr	56
Chapter 10 Flight Test Evaluation of Crew Workload by W.A. Wainwright	60
Chapter 11 Measurement of Aircrew Workload During Low-Level Flight by I.G. Lidderdale	69
Chapter 12 In-Flight Assessment of Workload Using Pilot Ratings and Heart Rate by A.H. Roscoe	78
Chapter 13 The Assessment of Workload in Helicopters by H.C. Muir and R. Elwell	83
Chapter 14 Assessing Pilot Workload For Minimum Crew Certification by J.J. Speyer, A. Fort, J.P. Fouillot and R.D. Blomberg	90
Chapter 15 Measurement of Pilot Workload by S.G. Hart	116
Chapter 16 Investigation of Workload Measuring Techniques: A Theoretical and Practical Framework by R.C. van de Grift	123
APPENDICES	131

CHAPTER I

INTRODUCTION

by

Alan H. Roscoe
Britannia Airways Limited
Luton Airport
Bedfordshire LU2 9ND
UK

1. DEFINING PILOT WORKLOAD

The term pilot workload is being used increasingly by those involved in the design and operation of modern aircraft. But a review of current literature on pilot workload makes it abundantly clear that there is still no generally acceptable definition of the term; nor is there any agreement on the best way of assessing it. Despite a large number of published papers and several seminars and workshops on the subject there has been little progress since the last Flight Mechanics Panel AGARDograph on pilot workload was published in February 1978. And yet in any discussion about pilot workload it is important — if not axiomatic — that there is a common understanding of what is meant by the term.

In the Introduction to that earlier AGARDograph (4) it was suggested that "it may be useful to consider workload as a multi-faceted concept, primary facets being formed by the three variables: demands of the flight task, pilot effort, and results. Minor or secondary facets can then be formed by the various methods used for assessing levels of workload. These will be largely dependent on the experience, discipline, and interest of the investigator". In 1982 O'Donnell (2) defined workload as "...an hypothetical construct which conveniently describes the interactions between multiple factors affecting the operator's response in an operational system". He went on to point out that "...such a broad and incomplete definition has value only if the factors underlying them can be identified, and if metrics to assess these factors can be specified". O'Donnell identified three broad categories of factors which contribute to workload, namely: taskload, operator variables, and response. Hart (3) referred to workload being a subjective experience resulting from a combination of several different dimensions; the three main dimensions being task-related, pilot-related, and outcome-related. Nineteen components of these main dimensions were suggested by Hart as being important in creating the total experience of workload. In a later paper Miller and Hart (4) referred to nine dimensions worth examining in detail when studying total workload: task difficulty, time pressure, own performance, mental effort, physical effort, frustration, stress, fatigue, and activity type.

The multidimensional nature of pilot workload has been accepted generally but with varying degrees of emphasis on the different aspects. For example, engineers concerned with predicting levels of workload for aircraft yet to fly tend to interpret workload as a set of demands (5)(6), although in this case the term 'taskload' would be more appropriate. And those investigators who measure performance as a means of assessing workload are inclined to emphasise the outcome-related aspect (7)(8).

Probably the most favoured interpretation of workload is pilot-related, usually in terms of effort. Using a questionnaire, Ellis and Roscoe (9) obtained the views of some 350 military and airline pilots and concluded that more than 80% of professional pilots think of workload in terms of effort. It is also an interpretation that agrees well with the influence on the piloting task of such individual factors as natural ability, training and experience together with physical fitness, age, and the idiosyncratic response to stress.

The individual nature of pilot workload led Ellis and Roscoe (9) to propose that a modified version of the definition used by Cooper and Harper in the introduction to their Handling Qualities Rating Scale (10) would be most appropriate, namely: Pilot workload is the integrated mental and physical effort required to satisfy the perceived demands of a specified flight task. There is evidence that the failure of a pilot to perceive the demands of a flight task correctly has been a causative factor in several accidents, and so the reference in the above definition to this aspect of workload reflects its importance. In discussing a conceptual framework for analysing workload Hart and Sheridan (11) refer to "...the operator's perception of what is required that is the proximate driving force behind the strategies selected and the resources committed, ...".

Of course, not everyone will agree with Ellis and Roscoe's definition of pilot workload but it is probably worth bearing in mind until presented with a more acceptable one. Other interpretations will be evident in later chapters.

2. THE NEED TO ASSESS PILOT WORKLOAD

Modern combat aircraft, with their increasingly complex systems and the need to fly faster and lower to avoid sophisticated defence systems, generate high levels of workload for their crews. But the level of workload must not be allowed to become too high if performance is not to suffer. Consequently there is a strong requirement to be able to assess workload at all stages in the design and development of these aircraft. This point was underlined by Milam (12) during a discussion on pilot workload in single-seat fighters when he stated: "Workload measurements, whether subjective or objective, should be available much earlier in the design process so that design options can be intelligently considered".

The introduction of flight management computers and improved autopilots into civil transport aircraft has tended to reduce the demands on the crew so that it is now necessary to optimise workload — rather than reduce it — to improve safety. And so, as with combat aircraft, it is important to evaluate the different aspects of workload at all stages of development (11)(13).

In the early design stages of projected systems, procedures, or aircraft it is most convenient to be able to predict levels of workload for different operational scenarios. Eventually, of course, such predictions will need to be verified by assessing workload in real flight (14)(15).

In the case of civil transport aircraft the findings of the President's Task Force on Crew Complement (16) have stimulated further effort into developing more reliable techniques for assessing workload — particularly in flight — in order to satisfy certification criteria for new aircraft (15)(17)(18)(19). The assessment of workload specifically related to crew complement certification is described in Chapters 3, 4, 10 and 14.

The influence of the Task Force findings has undoubtedly been largely responsible for the preponderance of descriptions of techniques and flight trials to assess workload in civil transport aircraft in this AGARDograph and in other reports. On the other hand, the technical difficulty of assessing workload in combat aircraft, together with the cautious interpretation of physiological responses necessitated by the possible effects of physical stressors, such as 'g', tend to discourage the use of similar techniques in military aviation.

For some years the evaluation of new or modified systems or operational procedures — especially those associated with the more demanding phases of flight — has often included some form of workload assessment (20)(21). Valuable experience in developing acceptable techniques has been obtained during flight trials, for example, to evaluate ski-jump take-offs (22), and low visibility approaches and landings (23).

3 TECHNIQUES FOR ASSESSING PILOT WORKLOAD

The search for reliable techniques for assessing pilot workload, especially ones that can be used in flight, has occupied a large number of researchers during the past decade or so. Various techniques have been examined in a multitude of experiments; in particular, the increased availability of general aviation trainers (GAT) in research laboratories has apparently encouraged a marked increase in the number of projects involving pilot workload. Unfortunately, of the many different techniques that have been proposed most are appropriate only for use in the carefully controlled conditions of the laboratory or flight simulator (1)(24)(25)(26).

Several criteria for workload assessment techniques have been proposed by various authors, they include sensitivity, diagnosticity, selectivity, intrusiveness, concordance, reliability, operator acceptance, and convenience (26)(27)(28)(29). Additionally, when assessing workload in aircraft the techniques must be compatible with flight safety (20). Whilst it might be reasonable to strive to satisfy many of these criteria in laboratory studies it would be impracticable to apply them too rigorously in the real world. For example, the need for increased sensitivity has been underlined (30), but variations between pilots, and even within the same pilot from time to time, may be greater than any small differences in workload detected by an unduly sensitive technique. As well as being unnecessary such a technique could well be a disadvantage when assessing workload in flight.

The increasing use of advanced autopilots and flight management computers has, especially in civil transport aircraft, caused a substantial decrease in the physical content of the total workload with a consequent relative increase in the cognitive or mental content. This change, which has been underlined by several authors, has added to the problem of assessing workload whatever techniques are employed (31)(11)(15).

The various techniques for assessing pilot workload can be classified loosely into three groups: objective, subjective and physiological.

3.1 Objective Techniques

These can be further divided into performance measures and analytic techniques.

3.1.1 Performance Measures

There is undoubtedly a relationship between workload and performance even though it may not be a simple one (11), but performance is not the only criterion — what it costs in terms of pilot effort and how likely is a pilot to become overloaded is of crucial importance. For instance a pilot may exert more effort and increase his workload as the demands on him increase to maintain performance. Conversely, as appears to happen more and more often to-day, the demands on him may be reduced and performance may suffer as the perceived workload becomes less due to complacency (32). A relationship of this kind precludes the use of performance alone as a reliable means of assessing workload. Nonetheless, it is important when assessing workload to define performance criteria and then to monitor the result. Instrumented aircraft and external measuring devices, such as kinethodolites sited on airfields to monitor approaches and landings, are ideal (17)(20). This is rarely possible but the use of video cameras to record crew activity and cockpit instrumentation is an alternative way of monitoring performance — used by several investigators (19), (see also Chapters 10, 13 and 16). Occasionally one might have to resort to harnessing the competitive instincts or desires for challenge, present in most pilots, to ensure performance at a reasonably optimum level.

In Chapter 14 Speyer and Fort describe a comprehensive performance criteria analysis technique used by Airbus to investigate the influence of new digital equipment to be installed in the A310. And performance measures form an integral part of an investigation into assessing pilot workload described by van de Graaff in Chapter 16.

To overcome some of the problems associated with measuring performance in the primary task to assess workload it is common practice in laboratory experiments to use some form of 'secondary' or 'loading' task (33)(34). Simply, the idea, based on the concept of spare capacity, is to compare levels of performance achieved on the 'loading' task alone with levels achieved when combined with the primary task. Various modifications to the basic technique have been made in an attempt to overcome many of the objections to the use of secondary task techniques in real-life situations (25)(35); but at present their use in flight does not seem to be all that practicable.

Shingledecker, in Chapter 2, discusses further the use of secondary task techniques; he also describes a novel version, the embedded secondary task, using radio communications, which seems to hold some promise for future use in flight. This technique should be appropriate for all three five minute flight tasks (appendices) although Shingledecker has applied it to the ILS approach as an example of its use.

Hart, in Chapter 13, as part of a battery of techniques, recommends the measurement of performance on a secondary task (time estimation) as well as on the primary task.

Secondary tasks are also employed as part of the methodology described by Donchin and Wickens in Chapter 6; and by Tole and Harris in Chapter 9.

3.1.2 Analytic Techniques

As mentioned earlier, many engineers and designers view workload in terms of the demands of the task. This is an interpretation of workload that supports the use of analytic techniques based on some form of time and motion study (5). Time-line analysis carried out in mockups, in flight simulators, or in real aircraft is used to build up a data store of physical activity associated with specific scenarios. From these data models can be constructed and indices of workload calculated, the taskload for a particular task or aircraft can be then predicted. As Milgram and his colleagues (36) observed: "The ability to analyse various aspects of crew activity during the carrying out of well defined flight scenarios is of great potential value, both as a developmental tool and for the design of flight decks and as an aid for ultimately complying with certification requirements".

Analytic techniques have been used by several airframe manufacturers to satisfy airworthiness requirements on crew complement — both from the ergonomic and from the workload points of view (17)(6).

Later in this volume (Chapter 3), Stone, Gulick, and Gabriel of the Douglas Aircraft Company describe in detail the use of task/time line analysis in the quantification of crew workload. The primary measure being the ratio of the time required for the task to the time available, within the constraints of a specific flight. An objective measure of workload sensitive enough to differentiate between alternative crew station layouts, displays, and controls is provided by a computerised technique based on comparative analysis. Special attention is given to those high workload procedures considered to be of special significance by designer.

The practical application of the methodology is demonstrated by reference to the MD-80 crew complement certification programme. In-flight collection of data together with the subsequent correlation analysis is of particular interest.

At Airbus Speyer and Fort used task/time analysis for the Static Taskload Analysis phase of a detailed programme of workload evaluation for the certification of the A310 for two pilot operation (see Chapter 14).

In Chapter 16 van de Graaff refers to the use of a video camera on the flight deck to monitor crew activity as well as to observe errors.

3.2 Subjective Techniques

Subjective reporting, in some form, by experienced test pilots is undoubtedly the most commonly used and probably the most reliable way of assessing workload in flight presently available. This observation should not be too surprising as, in many ways, subjective impressions of how hard a pilot is working — the amount of effort he has to exert to meet the demands of the task — are most relevant. As Hart (3) said "Workload is a subjective experience".

It has been suggested that subjective opinions are more reliable when a pilot is flying an aeroplane manually and Sanders (37) concluded that "the prospects of measuring mental load by subjective judgements are not high". On the other hand, Butterbaugh (13) wrote "subjective methods will continue to be valuable tools, especially because of the cockpit trend towards more monitoring and decision making tasks...". Subjective techniques must, of course, always be sensitive to preconceived ideas and bias, and evidence to this effect has been observed in experienced test pilots (15).

Various techniques for obtaining subjective opinions exist ranging from simple unstructured interviews and questionnaires for use after flight to sophisticated rating scales for use during flight. Post flight techniques usually have the advantage of simplicity and can provide valuable information on workload but they rely heavily on a pilot's ability to recall events and impressions that may have occurred some time previously. Nevertheless, unstructured or structured interviews and questionnaires are worthwhile and can be used with advantage to complement inflight measures (38)(39) (and Chapters 10 and 11).

In Chapter 4 Ruggiero and Fadden describe a Pilot Subjective Evaluation (PSE) technique for assessing workload which was used successfully during Boeing 767 Minimum Crew Size certification flight tests. The PSE, consisting of a post-flight questionnaire and a debriefing interview, was used to obtain information from both pilots for each test sortie. Data obtained in this way were used to validate time-line analysis and part-task simulator data in addition to providing final confirmation of workload levels experienced in this aeroplane.

As Hess (40) wrote "In all instances in which human opinion is elicited, there are definite advantages in obtaining quantitative responses". Well designed rating scales, used properly, provide a relatively inexpensive and convenient means of assessing pilot workload in a quantifiable form. The literature contains many references to different types of workload rating scales; indeed, it would seem that most people involved in the subjective evaluation of workload have designed their own scale!

The best known rating scale used in flight evaluation is the Cooper-Harper Handling Qualities scale (10), familiar to many test pilots and sometimes used — though mistakenly — to rate workload (41). The principle of the Cooper-Harper scale has been used as a model for several workload rating scales. For example, Wierwille (42) developed a modification of the scale called the Modified Cooper-Harper (MCH) which could be used for estimating pilot workload. This scale was subsequently compared with five other rating scales for assessing workload, it was concluded that the MCH was to be preferred and was, therefore, recommended for general use (43). Another ten-point rating scale based on the decision tree design of the Cooper-

Harper scale, and using the concept of spare capacity, has been developed with the help of practising test pilots at RAE Bedford (15). The use of this scale, which has already been used extensively for rating workload in flight (15)(22), is referred to in later chapters.

Also referred to later is a three dimensional rating scale known as the Subjective Workload Assessment Technique (SWAT). The three dimensions of workload: time pressure, mental effort, and psychological stress experienced are each rated on a three point scale. An overall rating of workload is calculated from a combination of the three individual ratings by the application of a conjoint scaling procedure. The technique requires a preliminary scale development for each subject pilot when the 27 possible combinations of ratings from the three dimensions are ranked (44)(45).

A flight simulator experiment to validate SWAT is described in Chapter 5 by Schick and Hann. The experimental task consisted of several 10 minute flights in an airport terminal area each ending in an approach and landing, the levels of difficulty being varied from flight to flight. SWAT ratings were obtained for each of the six segments into which the flight was divided.

Presumably, a similar application of SWAT could be used to assess workload in the 5 minute 'standard' flight task, the approach and landing, described in Appendix 1.

The SWAT technique and another interval scale (McDonnell Handling Qualities Rating Scale) are used together with pre- and post-flight ranking to assess pilot workload for eight different approach tasks in a study described in Chapter 15 by van de Graeff.

In Chapter 7 Biferno describes the use of a subjective rating scale which not only indicates 'what' the rating is but also 'why' it should be so.

Hart and her colleagues at NASA Ames (4) (46) have designed a set of bipolar rating scales incorporating nine dimensions related to workload. Although a single value for overall workload can be calculated the relative importance of each dimension can also be determined for each subject.

More recently Hart and her co-workers (47) have evaluated in flight a simpler scale having only five dimensions, stress, mental effort, fatigue, time pressure, and performance.

In Chapter 15 a new two-dimensional rating scale with six subscales is described by Hertz; this scale can be used to obtain an overall rating of workload. The importance of each of six factors is obtained by a simple pair-wise comparison.

Undoubtedly, the most appropriate time to assess workload is during flight and especially during the particular segment or task of interest (14)(15)(48). Consequently the value of a rating scale will increase if it is capable of being used during periods of high workload. There is some experimental evidence that subjective ratings given more than 15 to 30 minutes after the task are less reliable (49). The time period or segment of flight for which the rating applies may vary considerably and so to minimise the load on a pilot's memory it may be necessary to request ratings at frequent intervals.

It has been argued that trained test pilots (47) or pilots given special training (8) are necessary to use a rating scale efficiently. But Roscoe (15) has reported that a scale developed for the use of test pilots has been used successfully by airline pilots after only a brief introduction to the technique. The ease with which the scale was used was attributed to mention in the scale of spare capacity — a concept that seems to fit in well with pilots' ideas of workload.

In general, ratings do not disrupt the primary task but it has been suggested that during periods of high workload ratings may not be possible. Lidderdale (personal communication) was surprised to discover that pilots flying high speed low level sectors at night found it perfectly reasonable to give workload ratings on request, though on occasions the ratings were delayed by a minute or so due to preoccupation with the flying task. Other investigators have, likewise, reported delayed ratings during high workload phases but never complete omission (38)(39).

Rating individual components of workload may well be justified in a research environment but it is questionable whether the increased complication is worthwhile in practice. For example, it is difficult to imagine a pilot considering several dimensions of workload when asked to give an instantaneous rating during a particularly demanding flight task. As Stein and his colleagues (48) observed "..... this would make the workload response requirements more intrusive". Nevertheless, it is interesting to analyse the various constituents of total workload and also to determine their relative contributions during different flight tasks. However, one has to be careful in selecting possible components, for example, stress is certainly a part of workload — but the word stress is even more difficult to define than workload. And, because stress is used so frequently in common language it has several meanings some of which are outside any scientific context; yet stress is a component referred to in several multi-dimensional scales (44)(46)(47). On the other hand, the term pacing stress, or time stress — being much more specific — must be worthwhile identifying. As Hart and her co-authors (47) wrote on the subject of multi-dimensional scales: "One assumption that forms the basis of this approach is that individuals are able to assess the level of component variables more accurately and reliably than they can the combined experience termed 'workload'. This assumption may or may not be justified".

Workload rating scales are not always the prerogative of pilots, the same scale as that used by pilots or a different scale may be used by experienced observers to evaluate a pilot's workload. Two different scales, one a five point scale and the other a seven point scale, were used by pilots and observers respectively to evaluate workload in the Airbus A300FF workload trials (17). Later, during the A310 certification a seven point scale was used by both pilots and observers (see Chapter 14). British Aerospace used the same ten point scale for both pilots and observers during the flight evaluation of the BAe 146 (36)(Chapter 10). Although observers' assessments of workload must necessarily be incomplete, for instance mental activity can only be surmised, surprisingly good agreements between pilots' ratings and observers' ratings have been reported (17)(39).

3.3 Physiological Techniques

It is simple, convenient, and economical to assess workload by using a rating scale for pilots and, where practical, for observers. But the possibility of eliciting misleading data from inappropriate ratings lends support to the idea of using a second technique to augment subjective opinion (15)(50). The technique of measuring physiological variables to assess workload has been used for many years in a variety of situations. Physical effort, in particular, is easy to measure using such variables as heart rate and respiratory functions; however, the physical effort involved in piloting a modern aeroplane during normal manoeuvres is generally very low. But even though control forces are minimum, a pilot manually flying an aeroplane where precise and frequent control inputs may be required, as on landing, has a significant degree of neuromuscular involvement which can be detected by changes in a physiological variable such as heart rate (20). However, as the mental load, monitoring systems and making decisions, is becoming an increasing proportion of total workload — even in combat aircraft — physiological measures have to be selected carefully and used with great caution. Despite the many physiological variables studied in laboratories and simulator experiments on workload (26)(51), only a few are suitable for evaluating workload in flight.

As an increasing proportion of present day workload is mental, techniques that involve measuring brain activity must have an intuitive appeal. The electrical activity associated with brain functions can be recorded superficially as the electroencephalograph (EEG) and techniques such as the event related potential (ERP) based on this phenomenon have been developed specifically to determine mental load (52)(53)(54). As Donchin and his colleagues (50) pointed out: "The study of cognitive workload and of the allocation of processing resources to several tasks performed concurrently is, in fact, the area of research that has profited from the incorporation of ERP measures". Although not quite ready for routine use in aircraft some of the more advanced techniques using computer assisted analysis are worth considering for future use and two examples are described later.

Donchin and Wickens (Chapter 6) describe the practical application of a technique based on the Event-Related Brain Potential (ERP), and on the Sternberg Memory Search Task — two converging methodologies — to assess changes in workload during the 5 minute ILS approach and landing task defined in Appendix 1.

The problem of assessing operationally relevant mental workload is addressed in Chapter 7 by Biferno who, in the context of automated systems, considers pilot workload as language based mental activity. In addition to describing the use of a subjective rating scale for assessing mental workload Biferno also describes a technique being developed that uses a standard synthetic speech signal to elicit ERPs during flight.

The eyes obviously play a crucial role in flying an aeroplane whether the pilot is fully in the control loop or whether solely monitoring instruments. Different aspects of visual function have been suggested as being suitable variables for assessing workload (26). Measurements of eye movements or eye point of regard have been used in particular to estimate visual components of workload (55) but also as a means of estimating total workload (56). There is also some evidence that eye blinks may indicate changes in an individual's neurological state that are related to mental workload (57).

O'Donnell and Wilson (Chapter 8) consider that some physiological variables may be more specific than others and might, therefore, be used to complement each other. In referring to the Neuropsychophysiological Workload Test Battery (NWTB) currently being evaluated by the United States Air Force, it is suggested that transient cortical evoked responses and eye blink behaviour contribute complementary information about workload. These authors describe the practical use of these measures in assessing pilot workload during the approach and landing.

The psychophysiological techniques recommended by O'Donnell and Wilson should provide valuable data on the central information processing component of the piloting task as well as on the overall workload. But, whilst the technique shows promise and should be capable of further development for use in flight simulators, the application in real flight does not seem a practical proposition at present.

Tole and Harris (Chapter 9) discuss the measurement of eye point of regard to obtain information concerning workload during instrument flight. These authors suggest that the techniques of monitoring instrument scan patterns may be a potential candidate for workload assessment during the following:

- (a) Any situation in which instrument flight is required as part of the overall tasks.
- (b) Alterations in the design or layout of cockpit instrumentation.
- (c) Changes in controls which require visual monitoring.
- (d) Situations in which levels of fatigue may be unduly high.

Tole and Harris do underline the fact that instrument scan alone is not a complete indicator of workload, nevertheless, the technique may well be a useful complement when combined with others — for example, with a secondary task performance measure (Chapter 9).

The most used physiological variable for assessing workload in flight is heart rate (15)(39)(58)(59); it is easy and safe to use and is non-intrusive. Heart rate recorded and displayed in beat-to-beat, or instantaneous, form has the added benefit of demonstrating sinus arrhythmia which, especially in the absence of changes in mean heart rate, can be used to indicate changes in mental load (60)(61).

Heart rate is undoubtedly most valuable when used to augment subjective ratings of workload (15)(39)(47)(59), and the technique is described further in later chapters.

It is worth noting, though, that Hart (47) and Roscoe (15) have reported a greater level of reliability when using heart rate to support subjective ratings from handling pilots than from co-pilots.

The use of physiological variables to assess workload is based largely on the assumption that they reflect the level of neurological arousal determined by the demands of the flight task, i.e. by workload. It is important not to confuse emotion-induced arousal with task-induced arousal; unlike subjects taking part in laboratory experiments experienced professional pilots' heart rates, for example, are most unlikely to be influenced by emotional stressors during demanding flight (62)(63).

Several writers have criticised the use of physiological measures because of the lack of specificity or diagnosticity. Certainly, variables such as heart rate and sinus arrhythmia tend to be non-specific and to indicate only global workload — but often that is what is required in practice.

3.4 Combined Techniques

There is now strong evidence that a combination of different techniques provides the most reliable means of assessing workload in flight. Some form of subjective technique supported by a physiological measure appears to be a popular combination. Donchin, Kramer, and Wickens (50) considered that "there are circumstances in which subjective reports need augmentation, and in a subset of these circumstances ERPs may be very useful." A good correlation between heart rate and respiratory frequency, subjective ratings, overall performance, control activity, and model results has been reported by van de Graaff (58).

In Chapter 10 Wainwright describes the successful use of a battery of measures during a mini-airline flight trial to certificate the BAe 146 for two pilot operation. Subjective ratings from the two pilots and a flight observer during each sortie were complemented by post-flight questionnaires. The heart rates of both pilots were recorded continuously as a means of augmenting subjective ratings. Two video cameras situated on the flight deck recorded activity and performance. In part 2 of his chapter Wainwright suggests using a similar methodology for assessing workload during the hypothetical 5 minute approach and landing defined in Appendix 1.

Lidderdale (Chapter 11 Part 1) describes the low level flight trial of a combat aircraft in which crew workload was a most important issue. Workload assessments were obtained from a combination of in-flight subjective ratings using the Bedford scale, continuous recordings of heart rate from pilot and navigator, and post flight ratings using pairwise comparisons. A critical examination of the results of the two subjective techniques shows a high level of agreement thereby appearing to support strongly the use of in-flight ratings using the Bedford scale. In Part 2 of Chapter 11 Lidderdale suggests using the same technique — with some reservations about the use of an in-flight rating scale in a single seat aircraft — for the hypothetical 5 minute combat task (Appendix 2).

In Chapter 13, Muir and Elwell consider the implications of using a staged approach to the problem of analysing pilot workload in helicopters. They also describe the methodology — using a combination of subjective ratings, heart rate recordings, and video recordings, — to be used for assessing pilot workload during a forthcoming flight trial for the British Army. Clearly, this methodology may be applied directly to the hypothetical 5 minute helicopter flight task specified in Appendix 3.

For assessing workload during the 'standard' approach and landing task (Appendix 1) Roscoe (Chapter 12) recommends using in-flight ratings, obtained by means of the Bedford scale, augmented by recording pilots' heart rates.

Speyer and his colleagues at Airbus (Chapter 14) favour a combination of static and dynamic methods which include analytical techniques subjective ratings, performance measures, and heart rate.

In Chapter 15 Hart recommends the use of heart rate and sinus arrhythmia along with performance measures, task analysis and subjective ratings to improve the precision of workload evaluation.

van de Graaff in Chapter 16, describes a comprehensive experimental programme in which several techniques for assessing workload during different experimental landing approach tasks are evaluated. Heart rate is included in the battery of techniques along with pilot ratings, primary task measures (control activity, task performance, and error frequency), and model measures (control effort and decision load); in addition, time-motion parameter, secondary task performance, and crew-activity analysis using video recordings are used.

It should be noted that presently available techniques for assessing workload in flight do not result in absolute values. Rating scales and physiological variables are measures only of comparison; in other words, there has to be some form of standard or baseline whether defined in the experiment or as a function of a pilot's experience.

Because of the high cost of flying aeroplanes and the necessity eventually to assess workload in flight it is worth pointing out that although statistical probabilities are important they cannot be considered as definitive criteria for evaluating workload data.

4. SUMMARY AND GUIDANCE

The main purpose of this AGARDograph is to provide guidance for the reader who may wish to assess pilot workload in practical situations — rather than to be a comprehensive treatise on the subject. It has to be admitted, though, that at present workload cannot be assessed with any degree of precision or scientific certainty; nor is it likely that any significant improvement in the 'state of the art' will occur during the next decade. Nevertheless, cautious use of techniques selected from those described in the following chapters should provide valuable information on workload for designers and operators of aeroplanes as well as being of assistance in satisfying certification requirements.

Some techniques may be more appropriate than others for a particular requirement, for example, analytic techniques are more relevant during the design stage when attempting to predict workload. One advantage of using time-line analysis (see Chapters 3 and 14) is that the technique may be combined with some of the ergonomic studies associated with cockpit design.

An exercise of this type would be most appropriate in the design of advanced flight decks incorporating new systems. However, use of pre-flight analytic techniques can prove to be an expensive exercise and the anticipated overall cost might well be of concern.

During the last few years Airbus, Boeing, and McDonnell-Douglas have employed analytic techniques with notable success in the design and certification of new aircraft such as the A310, B757 and 767, and the MD-80. British Aerospace, on the other hand, did not consider it necessary to use pre-flight techniques when assessing workload on the BAe 146 with its more conventional flight deck. But, as Sulzer, Cor and Mohler wrote (14), "Final evidence of design adequacy is developed in flight tests because neither simulation nor analysis, without actual flight operations, can provide total substantiation that workload and crew duties are satisfactory when compared to existing operational aircraft."

Unfortunately, the choice of technique for use in flight is not a simple one. Whilst it is essential to monitor performance of the primary task when assessing workload the benefit of actual measurement as an assessment technique is less obvious. At Airbus, Speyer and Fort measured performance to compare workload levels for specific tasks (Chapter 14). Measurement of performance on secondary tasks has been suggested by other authors (see Chapters 2, 6, 15 and 16) but suitable techniques are not at present readily available for practical use in aircraft — although some do show promise.

From the view point of economy and ease of use, some form of subjective technique — a rating scale, perhaps complemented by a post-flight questionnaire — must be considered. If possible ratings should be capable of being given during flight without intruding into the piloting task. The Airbus and the Bedford scales (Chapter 10, 11, 12 and 14) are relatively simple and have been used successfully in flight on many occasions; but they result only in overall ratings of workload. Individual components of workload may be assessed by using the somewhat more complicated and sophisticated SWAT (Chapter 5) or the scale described by Hart in Chapter 15. Post flight questionnaires, such as those described by Rugeiro and Fadden (Chapter 4), by Wainwright (Chapter 10), and by Lidderdale (Chapter 11) can be used alone or in conjunction with other techniques and, being relatively simple to administer, have an undoubted value.

In view of the questionable reliability of subjective reporting of workload by pilots there seems to be a clear advantage in augmenting subjective data by means of an additional technique. There is increasing evidence that a number of physiological indices recorded from pilots may be used to complement their subjective assessments. In this respect, heart rate appears to be the most useful at present; it is safe, unobtrusive, and readily accepted by pilots. The technique, with modifications, has been used to assess workload in flight for over sixteen years by Roscoe (Chapter 12), and more recently by Wainwright, Lidderdale, and Speyer and his colleagues (Chapters 10, 11 and 14). Hart and van de Graaff (chapters 15 and 16) have also recorded pilots' heart rates in flight during experimental studies. Plots of beat-to-beat heart rate can be used not only to augment subjective ratings of workload over specific time periods but also to identify short term changes in workload that may not be readily apparent subjectively or by observation. In addition to using heart rate *per se* heart rate variability (sinus arrhythmia) can be of help in assessing mental workload for pilots engaged solely in monitoring.

Other physiological variables, eye movements, eye blinks, and, especially, evoked responses from the brain (Chapters 6, 7, 8, and 9) might well have a practical role to play in assessing workload with further development.

The prospective user is encouraged to select techniques suitable for his or her needs — further details being available from the relevant authors or from references cited in their chapters.

The table below summarises the main techniques featured in the different chapters.

CHAPTER	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
OBJECTIVE															
Performance measures (including video)									X			X	X	X	X
Secondary task techniques	X				X									X	X
Time-line analysis		X											X	X	
SUBJECTIVE															
Questionnaires			X						X	X					
Rating Scales				X		X			X	X	X	X	X	X	X
PHYSIOLOGICAL															
EEG (ERP)					X	X	X								
ECG Heart Rate (including HR variability)									X	X	X	X	X	X	X
Eye movements and blinks							X	X							

X used successfully in practice

REFERENCES

- 1 ROSCOE A H (Ed)
Assessing pilot workload. AGARDograph No.233, AGARD Paris 1978.
- 2 O'DONNELL R D
Historical foundation of the AFAMRL workload program. In: Proceedings of the Workshop on Flight Testing to Identify Pilot Workload and Pilot Dynamics.
- 3 HART S G
Theoretical basis for workload assessment research at NASA Ames Research Center. In: Proceedings of the Workshop on Flight Testing to Identify Pilot Workload and Pilot Dynamics AFFTC-TR-82-5, 1982.
- 4 MILLER R C
HART S G
Assessing the subjective workload of directional orientation tasks. In: Proceedings of 20th Annual Conference on Manual Control, 1984.
- 5 PARKS D L
Current workload methods and emerging challenges. In: Mental workload: its Theory and measurement, Moray N (Ed). 387-416, Plenum Press, New York 1979.
- 6 GULICK R
Validation of pilot workload estimates utilizing in-flight data. In: Proceedings of the Workshop on Flight Testing to Identify Pilot Workload and Pilot Dynamics. AFFTC-TR-82-5, 1982.
- 7 BRICTON C A
Pilot landing performance under high workload conditions. In: Conference Proceedings No.146 Simulation and Study of High Workload Operations AGARD Paris 1974.
- 8 HEFFLAY R K
CLEMENT W F
JEWELL W F
Overview of work in progress on non-intrusive assessment of pilot workload and pilot dynamics In: Proceedings of the Workshop on Flight Testing to Identify Pilot Workload and Pilot Dynamics. AFFTC-TR-82-5, 1982.
- 9 ELLIS G A
ROSCOE A H
The airline pilots view of flight deck workload: A preliminary study using a questionnaire. Royal Aircraft Establishment Technical Memorandum No FS(B)465, 1985.
- 10 COOPER G E
HARPER R P
The use of pilot rating in the evaluation of aircraft handling qualities. NASA Technical Note 5153 Washington DC, 1969.
- 11 HART S G
SHERIDAN T B
Pilot workload, performance and aircraft control automation. In: Conference Proceedings No.374 Human Factors Consideration in High Performance Aircraft AGARD Paris, 1984.
- 12 MILAM D W
A pilot's perspective of workload in single-seat fighters. In: Proceedings of the Workshop on Flight Testing to Identify Pilot Workload and Pilot Dynamics AFFTC-TR-82-5, 1982.
- 13 BUTTERBAUGH L C
Cockpit design for the future and challenge to workload measurement. In: Proceedings to the Workshop on Flight Testing to Identify Pilot Workload and Pilot Dynamics. AFFTC-TR-82-5, 1982.
- 14 SULZER R L
COX W J
MOHLER S R
Flight crew member workload evaluation. US DOT/FAA Report RD-82/21 Washington DC, 1981.
- 15 ROSCOE A H
Assessing pilot workload in flight. In: Conference Proceeding No.373 Flight Test Techniques AGARD Paris, 1984.
- 16 McLUCAS J L
DRINKWATER F J
LEAK H W
Report on the President's Task Force on Aircraft Crew Complement. Washington DC, 1981.
- 17 SPEYER J J
FORT A
Certification experience with methods for minimum crew demonstration. In: Conference Proceedings No.347 Flight Mechanics and System Design Lessons from Operational Experience AGARD Paris, 1983.
- 18 RUGGIERO F T
FADDEN D M
Pilot subjective evaluation: A practical supplement to traditional assessment of workload. In: Proceedings of International Conference on Cybernetics and Society IEE New York, 1982.
- 19 ANON
Flight test evaluation of workload. British Aerospace TSH 6273 Hatfield, 1983.
- 20 ROSCOE A H
Heart rate monitoring of pilots during steep gradient approaches. Aviat Space Environ Med 46 1410-1415, 1975.
- 21 HASBROOK H
RASMUSSEN P G
WILLIS D M
Pilot performance and heart rate during in-flight use of a compact instrument display. Report No.FAA-AM-75-12. FAA Office of Aviation Medicine. Washington DC, 1975.
- 22 ROSCOE A H
Handling qualities, workload, and heart rate. In: Survey of methods to assess workload Hartman B O and McKenzie R E (Eds). AGARDograph No.246 AGARD Paris, 1979.

- 23 ROSCOE A H
Pilot workload and economic Category 3 landings. In: Conference Pre-prints Aerospace Medical Association Annual Scientific Meeting, Washington DC, 1980.
- 24 WIERWILLE W W
WILLIGES R C
Survey and analysis of operator workload assessment techniques. Technical Report S-78-101 Naval Air Test Center Patuxent River, 1978.
- 25 WILLIGES R C
WIERWILLE W W
Behavioural measures of aircrew mental workload. Human Factors, 21, 549-574, 1979.
- 26 O'DONNELL R D
EGGEMEIR F T
Workload assessment methodology. In: Handbook of perception. Thomas A and Boff K (Eds) J Wiley and Sons, New York (In press).
- 27 JEX H R
Measuring aircrew workload: problems, progress, and promises. In: Proceedings of the Workshop on Flight Testing to Identify Pilot Workload and Pilot Dynamics. AFFTC-TR-82-5, 1982.
- 28 SHERIDAN T B
STASSEN H
Definitions, models, and measures of human workload. In: Mental workload: Its theory and measurement. Moray N (Ed), 219-234, Plenum Press, New York, 1979.
- 29 WICKENS C D
Engineering psychology and human performance. Merrill Columbus, Ohio 1984.
- 30 WIERWILLE W W
Determination of sensitive measures of pilot workload as a function of the types of piloting task. In: Proceedings of the Workshop on Flight Testing to Identify Pilot Workload and Pilot Dynamics. AFFTC-TR-82-5, 1982.
- 31 POPE A
BOWLES R L
A program for assessing pilot mental state in flight simulators. In: Proceedings of AIAA 20th Aerospace Science Meeting, Orlando, Florida, 1982.
- 32 ROSCOE A H
Pilot arousal during the approach and landing. In: Proceedings of the 32nd Annual Scientific Meeting of the International Academy of Aviation and Space Medicine. Madeira, Portugal (In press) 1984.
- 33 KNOWLES W B
Operator loading tasks. Human Factors, 5, 155-161, 1963.
- 34 OGDEN G D
LEVINE J M
EISNER E J
Measurement of workload by secondary tasks. Human Factors, 21, 529-548, 1979.
- 35 KELLY C R
WARGO M J
Cross-adaptive operator loading task. Human factors, 9, 395-404, 1967.
- 36 MILGRAM P
et al
Multi-crew model analytic assessment of decision-making demand and landing performance. In: Proceedings of 20th Annual Conference on Manual Control, Vol.II, 1984.
- 37 SANDERS A F
Some remarks on mental load. In: Mental load: Its theory and measurement. Moray N (Ed), 41-78. Plenum Press, New York 1979.
- 38 SPEYER J J
FORT A
Workload assessment for a 300FF certification. In: Proceedings of International Conference on Cybernetics and Society IEEE, New York, 1982.
- 39 WAINWRIGHT W A
BAe 146 - flight test evaluation of workload. Certificate Report HTD.R.460-00 SC0038. British Aerospace, Hatfield, 1982.
- 40 HESS R A
Non adjectival rating scales in human response experiments. Human Factors, 15, 275-280, 1973.
- 41 ELLIS G A
Subjective assessment pilot opinion measures. In: Assessing pilot workload. Roscoe A H (ed). AGARDograph No.233, AGARD Paris, 1978.
- 42 WIERWILLE W W
CASELI J G
A validated rating scale for global mental workload measurement applications. In: Proceedings of 27th Annual Meeting of the Human Factor Society, Vol.1, 1983.
- 43 WIERWILLE W W
SKIPPER J H
RIEGER C A
Decision tree rating scales for workload estimation. Theme and variations. In: Proceedings of 210th Annual Conference on Manual Control, 1984.
- 44 REID G B
et al
Development of multidimensional subjective measures of workload. In: Proceedings of International Conference on Cybernetics and Society. IEEE, New York, 1981.
- 45 REID G B
EGGEMEIR F T
SHINGLEDECKER C A
Subjective workload assessment techniques. In: Proceedings of the Workshop on Flight Testing to Identify Pilot Workload and Pilot Dynamics. AFFTC-TR-82-5, 1982.
- 46 HAUSER J R
CHILDRESS M E
HART S G
Rating consistency and component salience in subjective workload estimation. In: Proceedings of 18th Annual Conference on Manual Control, 1983.
- 47 HART S G
HAUSER J R
LESTER P T
In flight evaluation of four measures of pilot workload. In: Proceedings of the Human Factors Society 28th Annual Meeting, 1984.

- 48 STEIN E S
FABRY J
ROSENBERG B
Elusive goal of measuring pilot workload in general aviation. In: Proceedings of the Workshop on Flight Testing to Identify Pilot Workload and Pilot Dynamics. AFFTC-TR-82-5, 1982.
- 49 EGGE MEIR F T
CRABTREE M S
LaPOINTE P A
The effect of delayed report on subjective ratings of mental workload. In: Proceedings of the Human Factor Society 27th Annual Meeting, 1983.
- 50 DONCHIN E
KRAMER A F
WICKENS C D
Probing the cognitive infra structure with event-related brain potential. In: Proceedings of the Workshop on Flight Testing to Identify Pilot Workload and Pilot Dynamics. AFFTC-TR-82-5, 1982.
- 51 ROSCOE A H
Physiological methods. In: Assessing pilot workload. Roscoe A H (Ed) AGARDograph No.233, AGARD, Paris 1978.
- 52 ISREAL J B
et al
The event-related brain potential as an index of display monitoring workload. Human Factors, 22, 211-224, 1980.
- 53 WILSON G F
Steady state evoked potentials and subject performance in operational environments. In: Proceedings of the International Conference on Cybernetics in Society. IEEE, New York 1981.
- 54 KRAMER A F
WICKENS C D
DONCHIN E
Performance enhancements under dual-task conditions. In: Proceedings of the 20th Conference on Manual Control, 1984.
- 55 SIMMONS R
SANDERS M
KIMBELL K
Visual performance: A method to assess workload in the flight environment. In: Survey of methods to assess workload. Hartman B O and McKenzie R E (eds). AGARDograph No.246, AGARD, Paris 1979.
- 56 TOLE J R
et al
Visual scanning behaviour and mental workload in aircraft pilots. Aviation Space and Environ Med, 53, 54-61, 1982.
- 57 KIM W
ZANGEMEISTER W
STARK L
No fatigue effect on blink rate. Proceedings of 20th Anniversary Conference on Manual Control. NASA Conference Publication 2341 (11), 337-348, 1984.
- 58 van de GRAAFF R C
NLR research on pilot dynamics and workload. In: Proceedings of the workshop on Flight Testing to Identify Pilot Workload and Pilot Dynamics. AFFTC-TR-82-5, 1982.
- 59 SPEYER J J
et al
Heart rate monitoring analysis A310 flight campaign. Airbus Industrie/University of Paris. AI/V-F 304/83, 1983.
- 60 MULDER G
Sinus arrhythmia and mental workload. In: Mental workload: Its theory and measurement. Moray N (Ed), 327-344. Plenum Press, New York, 1979.
- 61 ROSCOE A H
Heart rate changes in test pilots. In: The study of heart rate variability. Kitney R I and Rompelman O (Eds). Clarendon Press, Oxford, 1980.
- 62 ROSCOE A H
Stress and workload in pilots. Aviation Space and Environment Med, 49, 630-636, 1978.
- 63 ROSCOE A H
Observations on heartrate changes in experienced pilots during flight. MD Thesis. Victoria University of Manchester, 1983.

CHAPTER 2

IN-FLIGHT WORKLOAD ASSESSMENT USING EMBEDDED SECONDARY RADIO COMMUNICATIONS TASKS

by

Clark A Shingledecker
Ergometrics Technology, Inc
4401 Dayton-Xenia Road
Dayton Ohio USA

INTRODUCTION

Traditional Secondary Task Measures

A widely accepted conceptual framework which forms the basis for many workload measurement techniques represents the human operator as a limited capacity information processing system. According to this general model, workload may be defined as the degree to which the operator's processing capacity is occupied by mental activities. Overload, and resulting performance decrement, occurs when capacity is insufficient to meet task demands. Since the momentary capacity of the operator is unknown and submaximal workload levels cannot be inferred from his or her performance on the task of interest, an indirect measure can be obtained by evaluating the amount of spare capacity available under a given set of task conditions.

The behavioral approach to assessing spare capacity involves the use of the secondary task technique. In this method, operators are given an additional information processing task to perform in conjunction with the task of interest. The rationale underlying the use of secondary tasks is that by applying an extra load which produces a total information processing demand that exceeds the operator's capacity, workload can be measured by observing the difference between single task and dual task performances. As noted by Ogden, Levine, and Eisaner (1), secondary tasks can be employed in two ways. Used as a loading technique, the method requires subjects to perform the secondary task under all circumstances with the intent of displaying overload effects in primary task performance. When secondary tasks are used as a workload measure, performance on the primary task is emphasized and secondary task performance is observed as an index of the workload of the primary task. Although specific research questions may require a choice of one of these applications, combined task decrement may also be used as an estimate of mutual interference and workload (2).

Unlike time-based analytical methods, the secondary task approach to assessing spare mental capacity has the potential for being sensitive to the degree of mental effort or attention devoted to information processing as well as to the temporal aspects of workload. The secondary task technique has the further advantage of producing a measure based on task performance, which is the variable that all workload measures ultimately must predict if they are to be of any value.

Although secondary task methodology has proven to be a useful technique for the investigation of cognitive processes, its practical application as a workload measurement tool has often been confined to the earliest stages of aircraft system design. As Schiffett (3) has noted, most workload measures have been developed for, and are most applicable to, the laboratory environment in which highly controlled, part task studies of workload can be conducted. When subsystems are combined to evaluate mission performance in the context of high fidelity simulations or flight tests, many workload assessment methods become difficult to employ because they are impractical or present potential safety hazards. As a result, workload measurement at the critical later stages of system development is often performed using relatively informal and qualitative techniques.

Three specific problems are encountered when traditional laboratory secondary tasks are considered for use during advanced development of aircraft. One practical problem is the physical instrumentation of the secondary task. In a flight test environment, and to a lesser extent in a simulator, introducing or adding any extra equipment to the crew station may be unacceptable. Even when sufficient space can be reserved, the possibility of obstruction or distraction caused by the additional instrumentation can limit the feasibility of using the secondary task.

A second problem with the implementation of secondary tasks is the possibility of intrusion on primary flight duties. Although some performance decrement may be tolerable, task interference can easily complicate the interpretation of data in test environments where measures of all performance variables may be unavailable. A more serious consequence of primary task intrusion in the flight test environment is the potential for compromising flight safety.

The final factor limiting the use of secondary task measures is operator acceptance (4). Whether used to induce stress or to measure reserve capacity, a secondary task is likely to produce misleading data if the operator fails to integrate it with his normal duties. Acceptance is a potential problem with all laboratory tasks because they are obvious, artificial additions to the crewstation and have little face validity or congruence with the general performance situation. Such test conditions can lead the operator to neglect the secondary task or, because of its novelty, allow it to assume an artificially high priority. Thus, lack of operator acceptance can become a major contributor to primary task intrusion as well as a source of measurement error.

Embedded Secondary Tasks

The embedded secondary task methodology was developed by Shingledecker et al (5) (6) to improve the practical utility of dual task measures for in-flight workload assessment, while retaining many of the scientific advantages associated with traditional laboratory secondary tasks. The concept of the embedded secondary task is based on the hypothesis that instrumentation limitations, task intrusion, and poor operator acceptance can be minimized by designing secondary tasks which are fully integrated with system hardware and with the crewmember's conception of the mission environment. By their nature, such tasks are realistic components of crewstation activity, yet their performance can be manipulated and measured independently of the primary activities of interest.

AD-P005-629

While several classes of aircrew activity are potential candidates for isolation and use as embedded tasks, radio communications tasks are particularly suitable for this purpose. The radio communications which are most useful as embedded tasks are those initiated by a message sent from another aircraft or a ground controller to a pilot whose workload is to be assessed. Upon detection and identification of a relevant message, the pilot must engage in a sequence of verbal responses and radio switching activities in order to meet the demands of the communicated request.

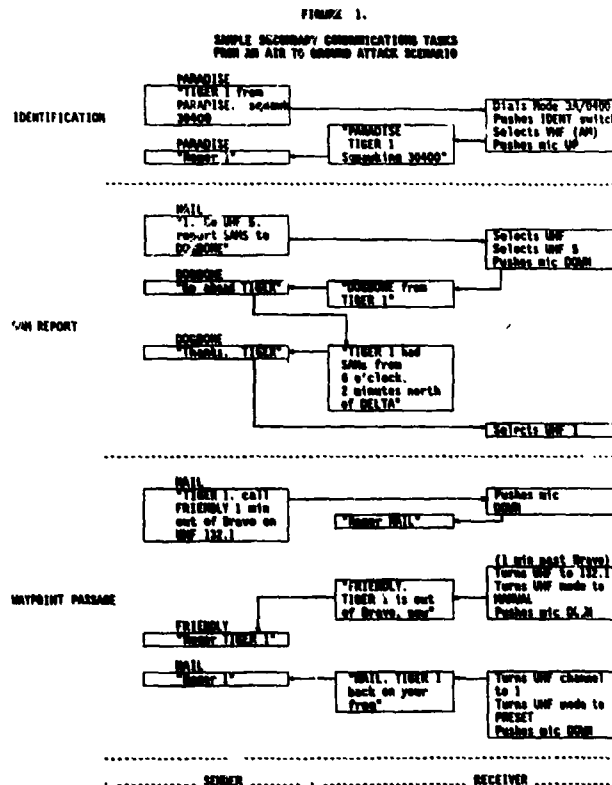
Such tasks closely resemble the nonadaptive discrete secondary tasks used in numerous workload studies and have many properties of good measurement tasks. Communications call upon a wide variety of information processing abilities and can be varied along several dimensions of complexity. Furthermore, no auxiliary crewstation equipment is necessary to control the experiment or to collect performance data. The opportunity for obstruction or peripheral interference is also minimized since the auditory channel is not shared by other tasks and verbal responses are generally unique to radio communications activities, while switch actions can be dealt with by the pilot's free hand. Most importantly, communications tasks are an integral part of a pilot's in-flight duties. As a result, lengthy training requirements are eliminated and high face validity is achieved. Additionally, the realistic nature of the activity makes artificial task interactions improbable because the pilot has predetermined priorities assigned to communications and other cockpit functions. These features make communications activities especially valuable for use as secondary tasks since pilots consider them to be important, but will normally devote less attention to communications as more crucial tasks become difficult to perform.

DESCRIPTION OF THE TECHNIQUE

Task Selection

The use of radio communications activities as embedded workload measures for high fidelity simulation or in-flight environments requires careful selection of the communications tasks to insure both realism and valid measurement. First, a group of candidate tasks must be identified which are relevant to the aircraft and mission of interest. Appropriate tasks may be obtained by interviewing operational pilots. In documenting these tasks, particular care should be taken to specify all verbiage used by the sender and receiver of the radio messages as well as the manual control actions required of the aircraft member. Additionally, the typical frequency and time of occurrence for each task should be noted.

Tasks which do not appear in the majority of interview responses or which vary in procedure among protocols should be eliminated from the group. Furthermore, those tasks which tend to take precedence over normal aircraft control functions should be avoided. For example, messages communicating threat would undoubtedly alter a pilot's normal attentional priorities and would shift any workload induced performance decrement to primary flight tasks. Some sample tasks which were obtained from single seat fighter attack pilot and which meet the requirements discussed above are shown in Figure 1.



Traditional discrete trial laboratory secondary tasks insure comparability of individual data points by repeating identical stimuli. Since communications tasks are not obviously comparable in their information processing demands, a second step that must be taken in task selection is workload scaling. Such scaling permits the experimenter to select a realistic combination of tasks for use in workload measurement which present equivalent estimated subsidiary loading levels. Shingledecker and his co-workers (4) evaluated three alternative apriori scaling techniques to achieve this purpose. Of the analytical and subjective methods which were tested, an information theoretical approach produced the highest correlation with dual task decrement scores.

This scaling technique is based on the assumption that the mental workload of communications tasks can be predicted by assessing the uncertainty associated with the reception of stimuli and execution of responses required of the pilot. Once a radio message is detected, the pilot must make two perceptual decisions to identify the intended receiver of the message and its sender. According to information theory, the demands associated with each decision can be estimated by determining the number of potential receivers and senders in the scenario and calculating a bit measure of the uncertainty of the decisions ($\log_2 N$). Thus, a message beginning with "Dogbone, this is Powder. . . ." would require the reception of 2.32 bits if there were five active receivers on the radio channel, plus one bit if there were two active message senders in the scenario.

Following these perceptual decisions, the pilot must make action decisions in response to the instructions received. Action decisions may require verbal and/or manual responses, and again may be quantified by determining the number of alternative actions that could be made. Thus, if a UHF radio channel change were required, the action sequence might involve the selection of a tuning mode with two alternatives (1 bit), turning a rotary control to one of twenty preset channels (4.32 bits) and pressing a microphone switch with two-positions to acknowledge the message (1 bit).

While verbal response decisions are more difficult to quantify in the information theoretic metric, a majority of these behaviors can be classified into one of two types. The simplest activity is a message confirmation which involves simple information conservation. Within this scaling method such responses are assigned a value of one bit. The second type of response requires the pilot to select a new receiver from among those active in the scenario and to report some information from cockpit displays or the external visual scene. In these cases the verbal response requirements are computed by summing the bits associated with selecting from among the available receivers, and adding a single bit for the report.

An overall estimate of the loading presented by a communications task is derived by summing the bit values calculated for all perceptual decisions and for each manual and verbal action decision in the task sequence. While this quasi-information theoretic method relies on assumptions of equiprobability of alternatives and independence of sequential actions, empirical tests indicate that it provides a reasonable estimate of secondary communications task loading. Values calculated for a set of candidate tasks may be used to select tasks with approximately equal load for workload assessment within a single flight scenario.

Workload Assessment

Once usable communications tasks are identified, their application for workload measurement closely follows the procedure normally used for traditional secondary tasks. Prior to testing the aircrew subjects should be briefed on the workload assessment procedure, emphasizing that their responses to some of the communications messages that will occur during the flight will be used to measure workload. They should be told to respond to these messages in a normal fashion, and to maintain primary flight task performance under all conditions (ie, the communications should not receive extra effort not afforded them in typical flying situations). Thus, they should respond to communications as quickly and accurately as possible, but not at the expense of primary flight control and management.

Prior to the test flight each participating pilot should review the communications tasks to be used for workload assessment. Finally, baseline single task performance should be recorded for each pilot on each of the tasks. This can be accomplished by presenting the tasks prior to take-off while the pilots are seated in the cockpits and are able to devote their full attention to the tasks. Performance scoring in both the single task baseline trials and in the in-flight test condition is accomplished by measuring each communication task completion time to the nearest 0.5 second. Times may be recorded manually beginning with the onset of the sender's message and ending with the final word of the pilot's response which completes the task sequence.

During the test flights the communications tasks should be presented to the pilots in accordance with a specified protocol developed to address the workload question of interest. Relative differences in workload between mission segments, cockpit design options etc are determined by comparing the magnitude of the difference between total task completion times for the baseline single task tests and the in-flight tests.

EXAMPLE OF USE

As in most other available workload measurement methods, the secondary communications task technique provides data which are interpreted in terms of comparisons among baseline conditions and various test conditions. Thus, no single example can address the potential range of workload questions or experimental design to which the technique is applicable. The example outlined below involves a hypothetical cockpit/system design issue. Equivalent examples could be developed to examine other comparative topics such as the impact on workload of flight experience, stressors or environmental conditions.

In the following case, the goal of the operational study will be to determine whether a new flight control system proposed for a twin jet transport aircraft reduces pilot workload during instrument approach and landing. It is assumed that previous test flights have revealed no objective evidence of major differences in flight performance between the current system and the proposed system. Two aircraft are available for the test, one equipped with the current flight control system and the other with the new system. Furthermore, five pilots who have equal flight time in the current system and have been thoroughly trained with the new system are available as test subjects.

Three types of communications tasks have been selected and scaled for use in the workload assessment. Each of these is initiated by air traffic control, but could be presented by an on board observer whose microphone is patched-in to the radios. The three messages are: 1) a request for radio frequency change (eg "FLYWAY 219, Contact approach on 118.1"), 2) a request to change transponder codes (eg "FLYWAY 219, Squawk 5133"), 3) a request for traffic information (eg "FLYWAY 219, do you see DELTA 1011?").

The pilots are briefed on appropriate response procedures and single task baseline performance is timed before the test flights. Each pilot flies the standard approach and landing twice in the current aircraft and twice in the aircraft equipped with the new flight control system. The four flights are accomplished in a randomized order determined for each pilot. Data from any approach and landing which does not meet the flight performance requirements specified in the experimental protocol are rejected and the trial is repeated.

The secondary communication tasks are relayed to the pilot according to a predetermined schedule starting with the initial transition to approach and ending with the touchdown. Six tasks (two of each type) are presented in addition to normal communications during the final five minutes of flight. Performance is scored by computing the time difference between baseline single task performance for each communication task and the performance during each occurrence of the task in flight. Mean decrement scores are computed for each task under the current and proposed flight control system and proposed flight control system conditions. A statistically significant reduction in decrement scores when using the new system would be interpreted as evidence for improved workload as a result of the design change.

LIMITATIONS

Like other operational test methods, the embedded secondary communications task technique can present problems of experimental control and precision of measurement which may affect the sensitivity of a workload assessment. Consequently, its value as a realistic methodology should not be allowed to outweigh the need for preliminary testing under part task simulation conditions. Both laboratory measurements and confirmatory flight tests are required to provide a complete and defensible workload analysis. Specific issues that should be considered when deciding to employ this method for flight test purposes include:

- 1 At present, no standardized secondary communication tasks are available for general use. Each application requires selection and scaling of tasks which are tailored to individual workload questions, specific systems and their missions.
- 2 The technique produces relatively few data points per unit time. Each task requires several seconds to perform and must occur with a relatively realistic frequency. As a result, embedded communication tasks are more suited to evaluating workload over extended periods of five or more minutes than to brief intervals of interest.
- 3 The method has not been tested to determine the degree to which different tasks produce diagnostic measures of workload. That is, it is not known which communications tasks are most sensitive to particular types of crew station loading. Available data indicate that communications tasks requiring manual activities (eg, radio tuning) tend to provide optimal measures of crew workload in tasks which involve aircraft control as a primary component.

REFERENCES

- 1 OGDEN G D Measurement of workload by secondary tasks. Human Factors, 21, 529-548 1979
LEVINE J M
EISNER E J
- 2 WICKENS C D Multiple resources, task-hemispheric integrity and individual differences in time-sharing.
MOUNTFORD S J Human Factors, 23, 211-229 1981
SCHREINER W
- 3 SCHIFFLETT S G An Annotated Bibliography, US Naval Air Test Center, SY-257R-76 1976
- 4 SHINGLEDECKER C A Subsidiary radio communications task for workload assessment in R&D simulations: I.
et al Task development and workload scaling. Air Force Aerospace Medical Research
Laboratory Technical Report, AFAMRL-TR-80-126 1980
- 5 SHINGLEDECKER C A Subsidiary radio communications tasks for workload assessment in R&D simulations: II
CRABTREE M S Task sensitivity evaluation. Air Force Aerospace Medical Research Laboratory Technical
Report, AFAMRL-TR-82-57 1982

CHAPTER 3

USE OF TASK TIMELINE ANALYSIS TO ASSESS CREW WORKLOAD

by

G Stone, R K Gulick and J F Gabriel
 Douglas Aircraft Company
 McDonnell-Douglas Corporation
 Long Beach, California, USA

INTRODUCTION

As systems have become more sophisticated, the role of humans in operating and maintaining them has grown more complex. There has been a steadily growing recognition that human characteristics, particularly limitations and abilities, must be considered in some depth in system design if design objectives are to be met.

The size and role of the crew represent critical design decisions. Mission performance has a direct relationship to the ability of the crew to carry out all of the required functions. If necessary functions overload the crew, some will be omitted and others ineffectively performed. If this is the case, automation may have to be considered. If the crew is underloaded, boredom and reduced performance may result, in addition to unnecessary costs being incurred. An additional crew member will increase weight, design costs, fuel expenditures, and training costs. It has been estimated that, for a commercial aircraft, an additional flight crew member can result in a 4 to 5 percent increase in direct operating costs. In the same manner, for a military aircraft fleet of 200 with a life-cycle of 20 years, costs can amount to several hundred million dollars for each additional crew member.

Issues of crew size were so critical in preliminary design work for proposals on antisubmarine warfare (ASW) and airborne warning and control system aircraft (AWACS) that Douglas Aircraft Company conducted research on the problem. The use of workload measures to assess the viability of a selected crew complement as well as other crew interfaces was considered. It was established that a workload assessment method should be capable of being applied early in the design phase, be expressed in quantitative terms, be understandable, and be relevant to the needs of the engineer. It must also have reasonable validity, be repeatable, be low cost, and need only a short turnaround time to produce results. Finally, the method must include consideration of the following: mission requirements and parameters, aircraft performance, equipment design, operational procedures, environmental factors, and crew station configuration.

The subject of workload has received extensive treatment in the literature (1 to 4) and is still being pursued in research and development efforts. Work is currently in progress throughout the industry on a number of varied approaches, including the following:

Subjective assessments employing rating scales.

Physiological measures, including heart rate variables, muscle activity or "arousal" indices, and more recently, electroencephalographic data such as the event-related potential

Performance and/or behavioral measures

Task/timeline analysis measures.

Of the items listed above, the task/timeline approach appeared to be the most easily implemented and could meet most of the established criteria. A model was developed by Douglas Aircraft Company to utilize this workload measure in the design, verification of design improvements, and certification of recent aircraft. This approach will be presented in this paper.

Task analysis may be defined as the systematic determination of the activities required of personnel in the performance of a function or set of functions. Workload analysis, which employs a task analysis base, provides an appraisal of crew task loading resulting from the sequential accumulation of task times. This permits an evaluation of the capability of the crew to perform all assigned tasks in the time allotted by mission constraints.

This analytic approach is derived from methods developed early in this century called "time and motion studies" which were aimed at making industrial workers more efficient in the performance of manual tasks. Task analysis was promoted as a useful tool in system design starting in the early 1950s.

In general, applications identified for task analysis include crew duty allocation and the assessment of design alternatives, personnel and training requirements, human reliability and safety, maintainability and workload. They are also used in the development of operational procedures. Several specific approaches have been developed (5).

In spite of certain limitations, the task/timeline methods seemed to offer promise for meeting many of our criteria such as quantitativeness, availability early in design and responsiveness to mission and operational parameters. It was equipment-oriented and met the needs of our designers. If applied consistently, it should be reliable.

Because there is no universally acceptable scale of workload, the data are normally used comparatively; that is, if a baseline workload were developed for an aircraft, or subsystems, or both, this could be used to determine if the system under consideration resulted in a greater, equal or less task workload than the baseline. In addition several configurations could be compared to determine which has the lowest workload and the percentage differences.

The task/timeline workload assessment methodology, when first applied in 1975, proved to be rather labor-intensive. It, however, showed promise of being suitable for computerization of many of the activities, ultimately resulting in reduced cost

AD-P005 630

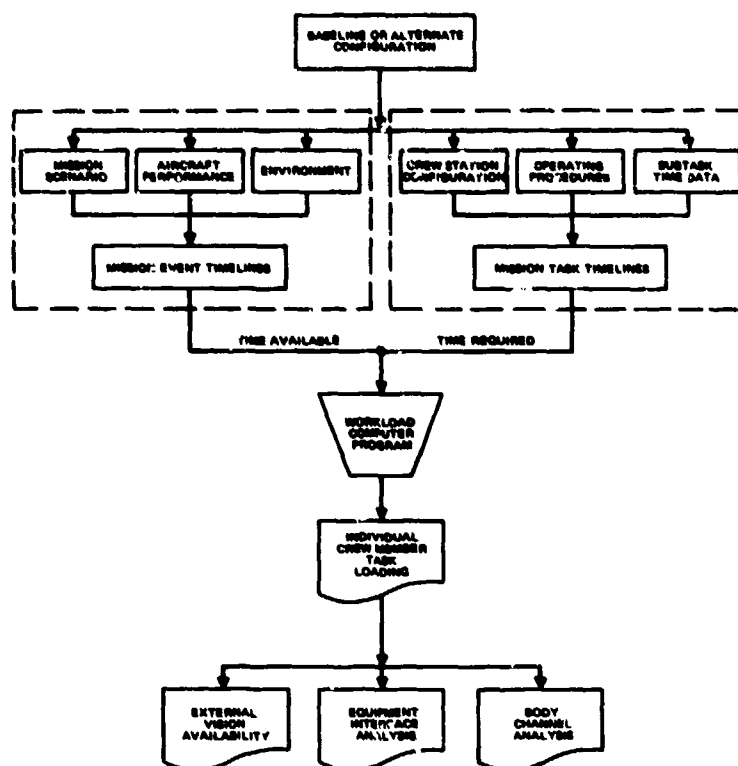


FIGURE 1. CREW STATION WORKLOAD ANALYSIS AND DESIGN SYSTEM (CRADS)

and time during the analysis process. Consequently, the task/timeline analysis approach was developed and partially applied to the DC-9-50 design. It has been used extensively in later design activities and is currently being used in flight deck and work station configuration development for Douglas Aircraft. It was applied to verify workload improvements for the MD-80 series and to demonstrate compliance with Federal Aviation regulations. For future aircraft now in design, it is employed in trade studies and for early design assurance that tasks during critical mission phases, including contingencies, can be performed by the available crew.

METHODOLOGY

Figure 1 shows the several analytic steps used in the basic approach to workload studies. Initially, mission analysis is employed to determine and size the parameters of the total functional system in which the crew and equipment will operate. The analysis is also used to organize the mission into phases and segments bounded by milestones to assist in system definition and establish top-level functions. This analysis is the foundation of an iterative descending hierarchy which, by further functional analysis and task analysis, ultimately reaches the irreducible task/subtask level (6).

The task analysis represents a detailed baseline that is effectively used to establish a comprehensive crew/equipment data store. At this level, comprehensive information on the tasks and task elements is developed from the previous mission and function analyses. The files of baseline data serve as the working library for preparation of crew workload reports.

WORKLOAD DEFINITION

Crew workload is defined as the ratio of time required by the crew to perform work tasks to the time available within a given mission, phase, or segment.

A workload index (WI) is computed which is expressed as the ratio of the total task performance time to the time available within the constraints imposed by mission requirements and aircraft design parameters. The basic formula for computing the index is:

$$WI = (T_R/T_A) \times 100$$

where T_R = time required
 T_A = time available

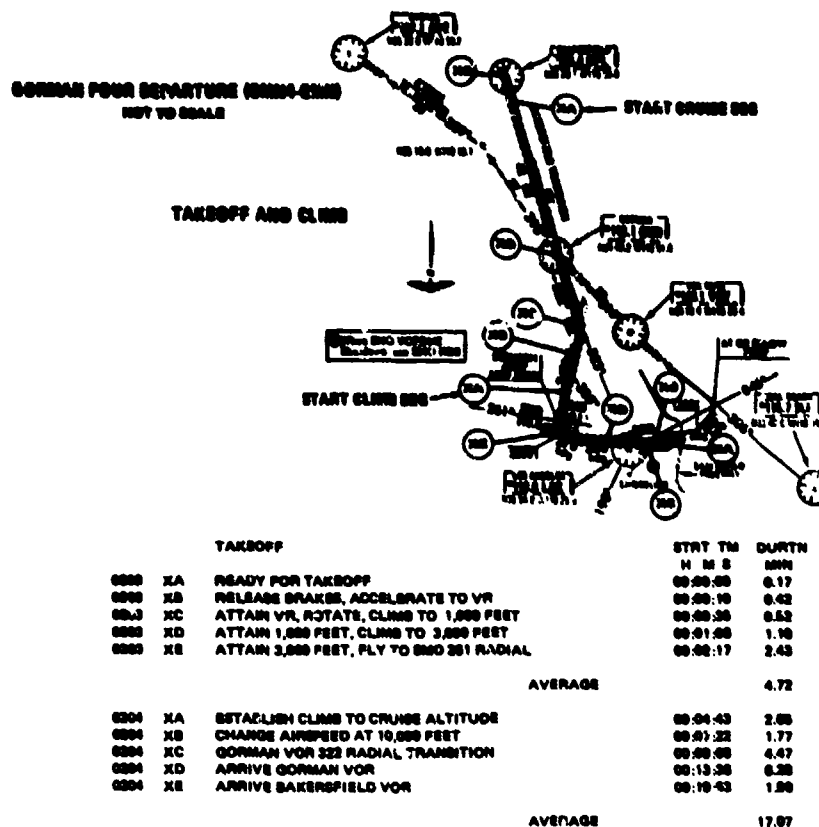


FIGURE 2. TAKEOFF AND CLIMB SEQUENCE

INPUTS

Time Available

To provide a framework for the detailed analysis, a scenario is divided into mission phases. Each phase is subdivided into discrete segments, bounded by specific operational milestones that define the start and end times based on aircraft performance characteristics, or mission parameters, or both.

Figure 2 illustrates the takeoff and climb phases at the start of a typical scenario from which time available parameters will be developed. The phases are then subdivided into segments — each bounded by a specific milestone (XA, XB, ..., XZ) denoting start and end times — which are derived from the aircraft performance characteristics and mission profile requirements. These relationships are shown in Figure 3. The difference between segment start and end times is the time available.

Time Required

Developing the time required begins with the use of crew station configuration drawings and proposed operating procedures for the aircraft and its specific equipment. All of the aircrew tasks and subtasks that must be performed between milestones are then detailed in chronological order and entered in the computer task file along with codes identifying specific equipment interfaces. The identity of the particular crew member performing that subtask and the specific body channels utilized (eyes, hands, etc) are also recorded. Working closely with flight personnel experienced in similar aircraft, a very detailed description of the procedures required to accomplish each mission segment is developed (down to a microlevel — eg move hand to switch). A typical sequencing is depicted in Figure 4.

As the detailed subtask and equipment listings are completed, individual "time required" values are assigned for each operator activity. These time estimates are derived from the following sources:

Index of Electronic Equipment Operability, developed by the American Institute for Research (AIR) (7)

A Douglas-developed model defining reach time as a function of distance.

Elect action time measurements recorded during procedural trials in a crew station development mockup.

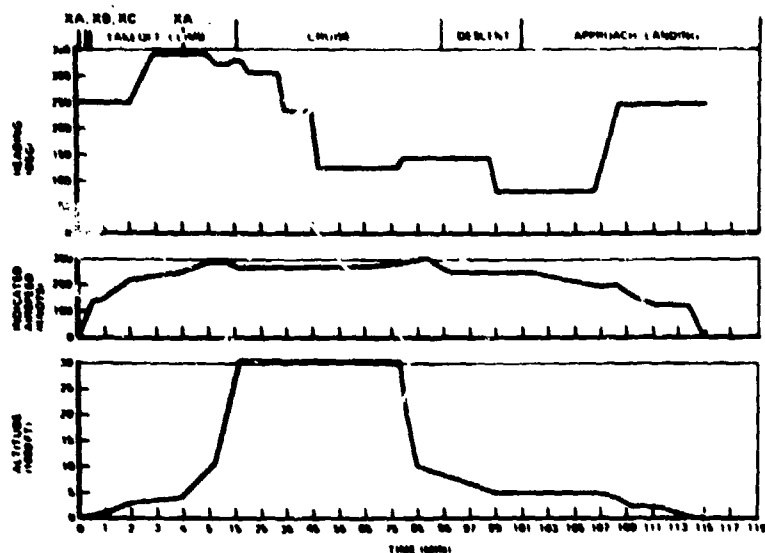


FIGURE 3. FLIGHT PROFILE RELATIONSHIPS - ALTITUDE, AIRSPEED, AND HEADING VERSUS TIME

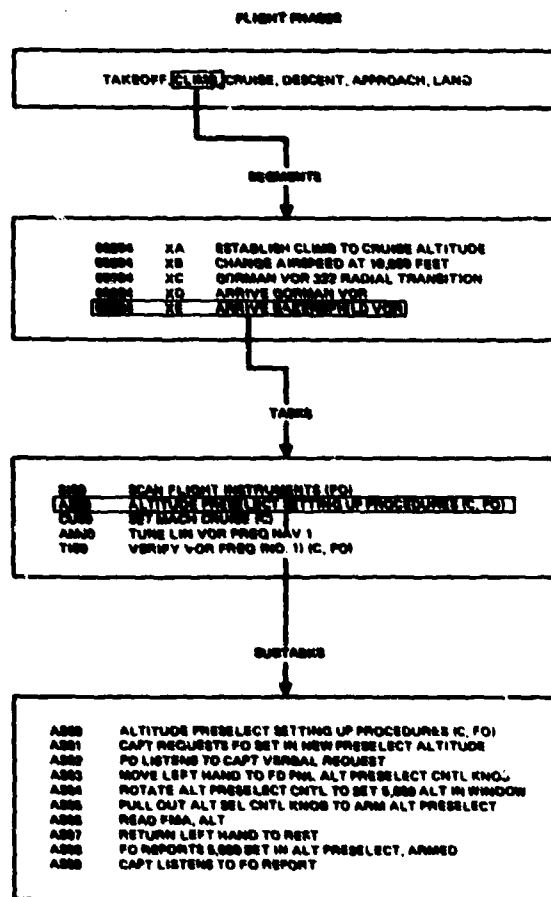


FIGURE 4. SEQUENCING STRUCTURE FOR COMPUTER INPUT

Time-referenced video recordings acquired during previous in-flight micromotion studies conducted by Douglas.
Timing verbal communications by stopwatch.

PROGRAM OUTPUTS

Equipment Interface Workload (Total Workload)

The crew workload produced by interfacing with equipment is defined as the total percentage of time that is utilized by the crew members in completing their assigned tasks while operating the aircraft during the mission. The computer program sums each individual crew member's task times and relates this to the time available in each segment of a particular mission. Since the program treats all subtasks as occurring in a series and does not reflect the human capability for simultaneous task performance such as listening while setting a switch, the workload values computed for an individual crew member can be considered conservative. These measures of workload are combined on a time-weighted basis to provide for an assessment of workload for each flight segment as well as an overall average for the entire flight. The program is capable of presenting both alphanumeric (Table 1) and graphic outputs (Figure 5) for further detailed analysis.

TABLE 1
CREW WORKLOAD INDEX SUMMARY
EQUIPMENT INTERFACE

AIRCRAFT: MD-8X		ANALYSIS: TEST		REVISION:	
FLIGHT FROM LAX		TO LAX			
FORM NO. 1	TITLE	START TIME H M S	END TIME H M S	WORKLOAD INDEX C P	
0001	XA READY FOR TAKEOFF	00:00:00	0.17	79.79	82.80
0001	XB RELEASE BRAKES, ACCELERATE TO VR	00:00:10	0.40	25.00	25.00
0001	XC ATTAIN VR, ROTATE, CLIMB TO 1000 FEET	00:00:30	0.30	47.79	47.79
0001	XD ATTAIN 1000 FEET, CLIMB TO 3000 FEET	00:01:00	1.10	39.39	47.69
0001	XE ATTAIN 3000 FEET, FLY TO 2ND 201 RADIAL	00:02:17	2.43	19.39	36.29
SEGMENT AVERAGE			4.72	30.02	36.72
0004	XA ESTABLISH CLIMB TO CRUISE ALTITUDE	00:04:03	2.05	0.00	12.00
0004	XB CHANGE AIRSPEED AT 10,000 FEET	00:07:00	1.77	17.79	40.00
0004	XC GORHAM VOR 325 RADIAL TRANSITION	00:09:00	0.07	10.00	10.00
0004	XD ARRIVE GORHAM VOR	00:13:36	0.00	10.00	10.00
0004	XE ARRIVE BAKERSFIELD VOR	00:19:53	1.90	20.00	42.01
SEGMENT AVERAGE			17.07	16.12	23.00
0301	XA CRUISE TO FRESNO VOR	00:21:07	11.33	7.00	4.01
0301	XB ARRIVE FRESNO VOR	00:33:07	12.00	0.07	0.01
0301	XC ARRIVE LINNEN VOR	00:46:00	0.00	0.07	0.00
0301	XD ARRIVE OAKLAND VOR	00:53:00	27.73	1.77	0.71
0301	XE ARRIVE AVAL VOR	01:10:04	10.50	47.00	30.01
SEGMENT AVERAGE			73.00	19.00	10.00
0401	XA VOR OVER FIM VOR, TURN, DESCEND	01:25:36	4.05	20.00	20.00
0401	XB ARRIVE SADDLE INTERSECTION	01:29:51	0.57	10.10	0.00
0401	XC ARRIVE 5000 FEET	01:40:25	2.03	20.00	20.00
SEGMENT AVERAGE			7.00	20.00	20.10
0403	XA VOR 5000 VOR, TURN TO 045 DEGREES	01:43:15	4.07	10.70	20.07
0403	XB TURN TO 275 DEGREE	01:47:20	1.40	0.01	0.00
0403	XC INTERCEPT ILAS LOCALIZER	01:49:17	1.40	10.17	10.00
0403	XD ARRIVE 2200 FEET	01:50:40	1.70	10.07	20.10
0403	XE ARRIVE OUTER MARKER FLY TO MIDDLE MARKER	01:50:34	2.07	10.71	21.00
SEGMENT AVERAGE			11.90	20.70	20.01
0404	XA CONTINUE DESCENT, MIDDLE MARKER TO MAIN GEAR TOUCHDOWN	01:50:14	0.30	20.37	20.00
0404	XB HALLOUT FROM TOUCHDOWN TO RUNWAY CLEARANCE	01:50:33	0.00	40.70	10.71
SEGMENT AVERAGE			1.10	27.97	20.00
OVERALL AVERAGE			116.35	17.01	19.50

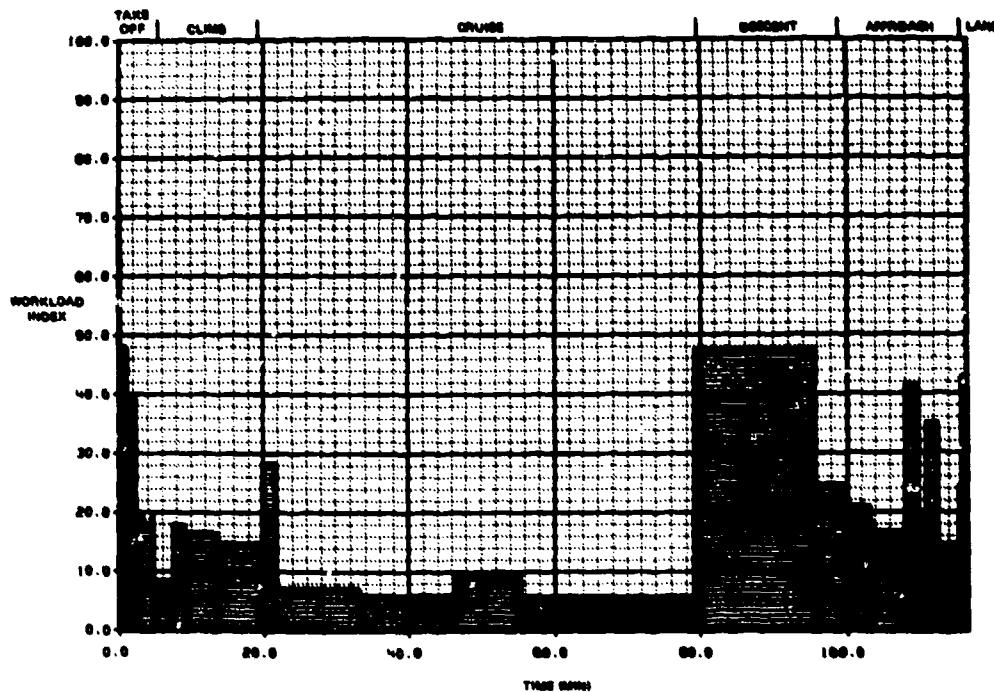


FIGURE 5. EQUIPMENT INTERFACE CREW WORKLOAD - CAPTAIN

Body Channel Workload

The quantification and evaluation of flight crew workload involves consideration of the overt physical actions taken by the flight crew to operate the aircraft. The program then determines the detailed work allocation as a five-channel input/output subsystem on a task-time basis for each crew member. It reflects a composite of the physical actions, reactions and perceptions necessary to fly an aircraft along a prescribed flight path. The flight crew workload analysis thus produces results in tabular and graphic format, reflecting the combined duty cycle of total visual, aural, vocal, and body extremity activity.

All flight crew subtasks are coded in accordance with the following body channel scheme:

- V/A — Verbal/aural tasks
- IV — Internal visual tasks
- L — Left-hand tasks
- R — Right-hand tasks
- F — Foot tasks

The overall flight deck activities involved in each flight segment are then analyzed in terms of the individual body channel utilization as a ratio of time required to time available. The results enable specific deficiencies to be identified in the functional arrangement of equipment through examination of peak values that might cause crew overload for an individual body channel.

Examples of the alphanumeric and graphic outputs are shown in Table 2 and Figure 6, respectively.

External Vision Availability

Time is required for crew members to view cockpit displays and controls during the course of the flight, and the remaining time can be considered as available for crew members to scan the outside environment. This analysis determines the amount of time available for a crew member to scan the airspace for traffic as well as to keep the runway in view during operations in the terminal area, both of which are important duties from a safety viewpoint.

The computer program examines data in the vision task file, sorts the data, and prints out the external vision time available for crew members as a function of the milestone start times and duration. In addition, for a two-pilot aircraft, a routine is provided to combine the Captain's and First Officer's external viewing time and present the information in graphic form so that total external vision available to both crew members may be ascertained throughout the flight. Typical vision analysis data outputs are shown in Table 3 and Figure 7.

Additional Capabilities

The amount of detailed information coded in the data files of the workload program provides additional analytic capability. The following crew interface relationships can also be evaluated:

TABLE 2
BODY CHANNEL UTILIZATION SUMMARY - CAPTAIN

AIRCRAFT: MD-XX ANALYSIS: TEST REVISION:
FLIGHT FROM LAX TO LAX

CREW MEMBER = C:CAPTAIN

FUNC M,ST TITLE

FUNC	M,ST	TITLE	STRT	TH	GIRTH	BODY CHANNEL INDEX					
						V/A	IV	L	R	F	
0203	KA	READY FOR TAKEOFF	00:00:00	0.17	30.6	32.7	0.0	15.7	0.0	0.0	
0203	KB	RELEASE BRAKES, ACCELERATE TO VR	00:00:10	0.42	9.6	0.0	0.0	2.4	0.0	0.0	
0203	KC	ATTAIN VR, ROTATE, CLIMB TO 1000 FEET	00:00:35	0.32	15.5	20.5	16.1	21.0	0.0	0.0	
0203	KD	ATTAIN 1000 FEET, CLIMB TO 3000 FEET	00:01:06	1.10	20.0	17.6	0.0	9.0	0.0	0.0	
0203	KE	ATTAIN 3000 FEET, FLY TO 2ND 261 RADIAL	00:02:17	2.43	6.7	12.0	0.0	6.1	0.0	0.0	
AVERAGE					4.72	12.37	10.35	1.77	0.75	0.71	
0204	KA	ESTABLISH CLIMB TO CRUISE ALTITUDE	00:04:43	2.65	2.9	3.1	0.0	2.3	0.0	0.0	
0204	KB	CHANGE AIRSPEED AT 10,000 FEET	00:07:22	1.77	9.7	7.6	0.0	2.0	0.0	0.0	
0204	KC	CORRMAN VOR 122 RADIAL TRANSITION	00:09:08	4.47	9.1	7.2	0.0	2.0	0.0	0.0	
0204	KD	ARRIVE GORDMAN VOR	00:13:36	6.28	6.1	9.9	1.0	6.0	0.0	0.0	
0204	KE	ARRIVE BAKERSFIELD VOR	00:19:33	1.90	13.9	13.9	0.0	7.4	0.0	0.0	
AVERAGE					17.07	6.90	0.67	0.66	0.43	0.0	
0301	KA	CRUISE TO FRESNO VOR	00:21:47	11.33	1.3	5.4	0.0	3.2	0.0	0.0	
0301	KB	ARRIVE FRESNO VOR	00:33:07	13.22	1.3	4.6	0.0	2.6	0.0	0.0	
0301	KC	ARRIVE LINDEN VOR	00:46:20	9.09	1.9	7.3	0.0	4.3	0.0	0.0	
0301	KD	ARRIVE OAKLAND VOR	00:55:20	21.73	2.1	3.5	0.0	1.6	0.0	0.0	
0301	KI	ARRIVE AVENAL VOR	01:19:04	16.53	22.9	23.9	4.5	9.0	0.0	0.0	
AVERAGE					73.82	6.30	9.03	1.01	3.31	0.0	
0401	KA	VIR OVER FIN VOR, 1UPN, DESCEND	01:35:36	4.25	9.6	14.0	0.0	6.1	0.0	0.0	
0401	KB	ARRIVE SADDLE INTERSECTION	01:38:51	0.57	0.0	16.2	0.0	0.0	0.0	0.0	
0401	KC	ARRIVE 5000 FEET	01:40:25	2.83	10.0	6.1	0.0	0.0	0.0	0.0	
AVERAGE					7.65	10.85	11.67	0.0	4.40	0.0	
0403	KA	VIR 2ND VOR, TURN TO 068 DEGREES	01:43:15	4.67	6.3	12.0	0.0	5.6	0.0	0.0	
0403	KB	TURN TO 225 DEGREES	01:47:55	1.70	20.4	19.7	0.0	0.9	0.0	0.0	
0403	KC	INTERCEPT ILAX LOCALIZER	01:49:37	1.20	6.4	11.9	0.0	7.1	0.0	0.0	
0403	KD	ARRIVE 2200 FEET	01:50:49	1.75	11.7	22.2	0.0	5.3	0.0	0.0	
0403	KE	ARRIVE OUTER MARKER FLY TO MIDDLE MARKER	01:52:34	2.67	6.6	7.6	0.0	1.3	0.3	0.3	
AVERAGE					11.90	8.40	13.99	0.0	5.25	0.11	
0404	KA	CONTINUE DESCENT, MIDDLE MARKER TO MAIN GEAR TOUCHDOWN	01:55:14	0.32	23.1	2.0	2.9	2.6	2.6	0.0	
0404	KB	ROLLOUT FROM TOUCHDOWN TO RUNWAY CLEARANCE	01:55:33	0.00	5.4	6.4	12.0	12.0	0.3	0.3	
AVERAGE					1.12	13.45	6.21	9.72	9.93	6.72	
OVERALL AVERAGE					116.35	7.34	9.01	0.96	4.00	0.10	

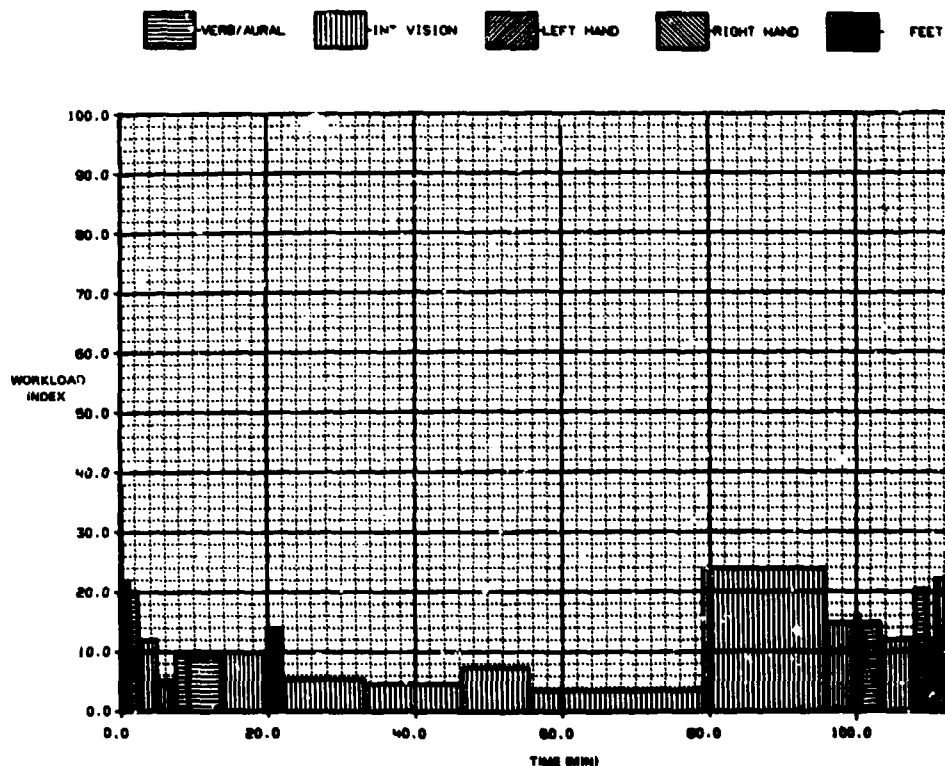


FIGURE 6. PEAK CHANNEL USAGE - CAPTAIN

TABLE 3
EXTERNAL VISION AVAILABILITY ANALYSIS

CREW MEMBER : CAPTAIN				STRT TM		DURTH		EVA INDEX	
PUNC PLST	TITLE	H	M	S	NIN	IV	EV		
00401	XA ARRIVE AT 10 MINUTE WARNING MPT	00	00	00	5.00	4.07	95.1		
00401	XB INITIATE SLOWDOWN	00	05	00	3.75	31.49	68.5		
00401	XC ARRIVE AT DROP ALTITUDE	00	08	45	1.25	19.35	80.7		
00401	XD ARRIVE AT CARP	00	10	00	0.50	29.07	70.9		
00401	XE ACCELERATE TO 350 KIAS - START DESCENT	00	10	30	1.30	33.72	64.3		
00401	XF LEVEL OFF AT 300 FEET	00	11	40	1.00	19.13	80.9		
AVERAGE						12.00	19.27	80.73	
OVERALL AVERAGE						12.00	19.27	80.73	

CREW MEMBER : FIRST OFFICER				STRT TM		DURTH		EVA INDEX	
PUNC PLST	TITLE	H	M	S	NIN	IV	EV		
00401	XA ARRIVE AT 10 MINUTE WARNING MPT	00	00	00	5.00	11.01	80.2		
00401	XB INITIATE SLOWDOWN	00	05	00	3.75	14.19	83.8		
00401	XC ARRIVE AT DROP ALTITUDE	00	08	45	1.25	24.88	75.1		
00401	XD ARRIVE AT CARP	00	10	00	0.50	36.30	43.5		
00401	XE ACCELERATE TO 350 KIAS - START DESCENT	00	10	30	1.30	20.87	71.1		
00401	XF LEVEL OFF AT 300 FEET	00	11	40	1.00	11.57	88.6		
AVERAGE						12.00	17.03	82.97	
OVERALL AVERAGE						12.00	17.03	82.97	

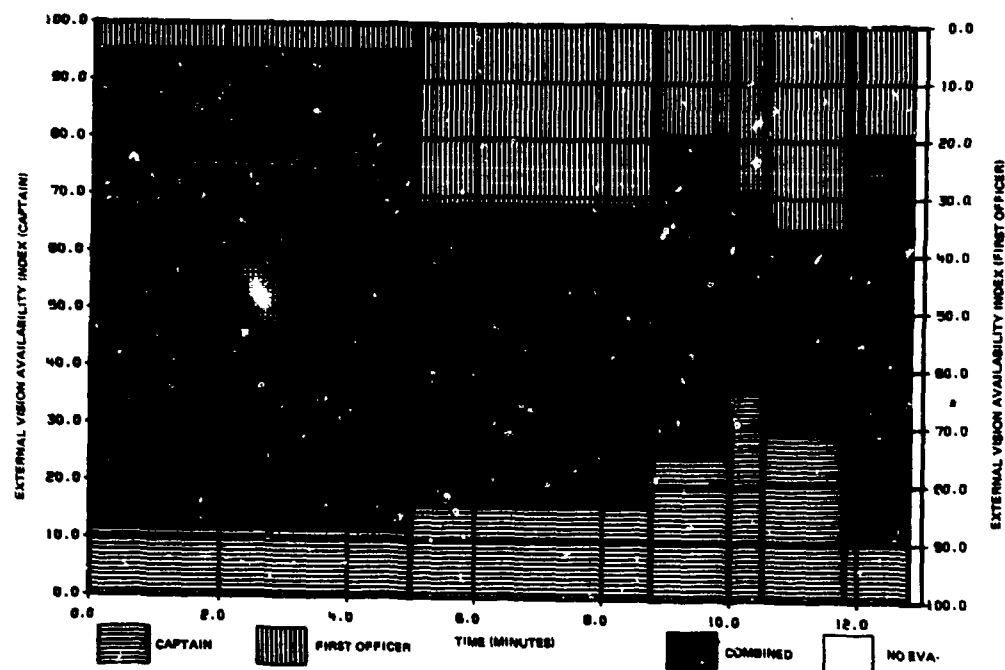


FIGURE 7. COMPOSITE EXTERNAL VISION AVAILABILITY

- 1 each system;
 - 2 each piece of equipment by part number;
 - 3 controls and displays and
 - 4 the effect of their location based on frequency of use.
- In addition, the responsible design groups can be identified.

These measures can be employed to evaluate crew work stations in preliminary design, including the proposed system control and display layouts, operational procedures, and to aid in the certification of new aircraft by validating the design as it applies to the man/machine interface. Additionally, various configurations can be examined in normal operational and in degraded modes where equipment failures have occurred. This latter capability is of great value as it allows analysis of conditions in which the workload may be such as to jeopardize mission accomplishment or safety.

VALIDATION

Because the crew workload index is a function of the ratio of the time required (T_R) to the time available (T_A), there are two aspects to be validated: 1. the segment times which are based on aircraft performance and establish the time available, eg brake release to aircraft rotational velocity (T_A), and 2. the time required (T_R) to perform the tasks within each segment.

The aircraft performance data used to develop the phase and segment times in the flight profile were provided by the Aerodynamics group of Douglas Aircraft, and were validated during engineering test flights. Therefore, they do not require further substantiation. The tasks and task sequences, jointly developed by Human Factors Engineering and Flight Operations, contain all cockpit interface activities considered necessary for effective and safe completion of the flight scenario. These interface activities were verified using a fixed base mockup. Validation of computed task times was therefore needed to ensure that they correspond realistically to actual in-flight times. The methodology for validating the data base task times is described in the following text.

Three flight test programs were conducted to collect data to be used in the validation process. The first set of data was collected during the certification flight of the DC-9-50 in approximately 1977. As part of the validation, a dedicated flight test was conducted that duplicated the scenario used in the MD-80 analytic workload study. This provided timeline data as well as verification of procedures used in the analysis. In addition, during the MD-80 crew complement certification process, a series of test flights was conducted in the high density US Eastern Corridor under airline operating conditions to satisfy Federal Aviation Regulations concerned with the minimum flight crew required for safe aircraft operation. There were nine consecutive days of flying, a total of 55 separate legs with a crew of three two-man teams, each composed of an FAA pilot and a Douglas pilot. Videotapes of flight deck activities recorded during these flights were studied using a micromotion analysis technique to obtain in-flight task time data. Some 122 tasks were examined with relevant human performance times tabulated.

A sample frame of the video tape, shown in Figure 8, indicates the units in which the tasks can be time ie, hours, minutes seconds, and tenths of a second. This is accomplished with a digital time generator which superimposes these data directly on the video tape (eg, 3 hours, 25 minutes 36.3 seconds). On the actual tape, the resolution is sufficient to distinguish individual controls and displays, allowing for precise determination of physical motion times.

Table 4 presents an example of three tasks and their comparative crew workload data base and in-flight measured times. In all, 122 tasks were examined in this manner. The results are shown in Figure 9 illustrating the linear regression line of the 122 points. An excellent correlation was obtained with a coefficient equal to 0.81.

As a result it was concluded that the task/timeline analysis procedure provides a reasonably accurate index for predicting the time required to complete observable tasks within the constraints of an actual mission. The detailed methodology and results of the data base validation process are presented in a previous report (8).

APPLICATIONS

Aircraft Comparison During Early Design

The comparative analysis capabilities of the program enable the new design to be compared to an existing aircraft that is known to have an acceptable workload profile and is duly certified. The existing aircraft will be referred to as the MD-X. The



FIGURE 8. SAMPLE FRAME FROM VIDEOTAPE - IN-FLIGHT RECORDING

TABLE 4
IN-FLIGHT AND DATA BASE
TASK COMPARISON

IN-FLIGHT	TIME (SEC)	DATA BASE	TIME (SEC)
ADJUST HEADING KNOB, PUSH ILS BUTTON			
CAPTAIN MOVES HAND TO HDG SEL KNOB FROM REST, ADJUSTS KNOB, MOVES HAND TO ILS BUTTON - PUSHES - RETURNS HAND TO REST		a. CAPTAIN REACHES TO HDG SEL KNOB	0.83
		b. ROTATES TO SET HEADING IN WINDOW	3.83
		c. MOVES HAND TO ILS BUTTON	0.36
		d. PUSHES BUTTON	0.57
		e. VERIFIES BUTTON ILLUMINATES	0.20
		f. RETURNS HAND TO REST	0.60
	6.4		6.07
SET RADIO ALTIMETER			
FIRST OFFICER SETS NO. 2 RADIO ALTIMETER WITH RIGHT HAND (REACHES AND RETURNS TO REST)		a. FIRST OFFICER MOVES HAND FROM REST TO NO. 2 RADIO ALTIMETER KNOB	0.84
		b. ROTATES TO SET BAROMETER	1.30
		c. RETURNS HAND TO REST	0.84
	2.8		2.98
SET ILS FREQUENCY - NAV 1 AND 2			
TIMED FROM FIRST OFFICER'S HAND ON NAV 2 - SETS FREQ, REACHES TO NAV 1 FREQ KNOB, ROTATES TO SET, RETURNS HAND TO REST		a. FIRST OFFICER SETS NAV 2 IN WINDOW	3.29
		b. MOVES HAND TO NAV 1 FREQ KNOB	0.86
		c. ROTATES TO SET FREQ IN WINDOW	3.29
		d. RETURNS HAND TO REST	0.72
	6.5		7.26

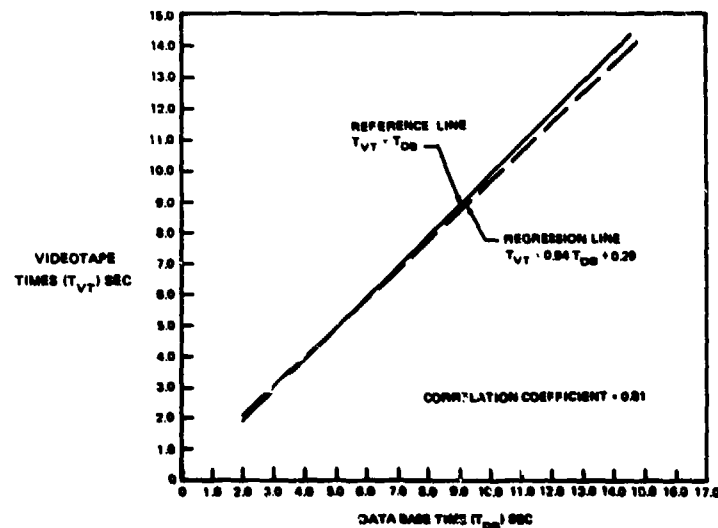


FIGURE 9. TASK TIME VALIDATION - VIDEOTAPE VERSUS DATA BASE TIMES (122 TASKS)

configuration incorporates a digital flight guidance system and autothrottle/autopilot capabilities. It also features conventional instrumentation displays. The new aircraft, designated the MD-XX, is equipped with a flight management system integrated with an automatic flight control system. Four electronic (CRT) instrument displays feature redundant primary flight and navigation displays, while two multifunction displays incorporate such features as phase-of-flight display, caution/warning alerts, fault/limit lists, and procedure/checklists.

In this example, the two aircraft are compared using a flight scenario involving the critical phases of descent, approach, and landing at LaGuardia airport in New York. The results of this analysis are shown in Figure 10 which illustrates the workloads of the Captain and First Officer. It is significant to note that while the operational systems of the advanced flight deck are sophisticated, there appears to be only a slight difference in workload compared to the baseline aircraft. While the First

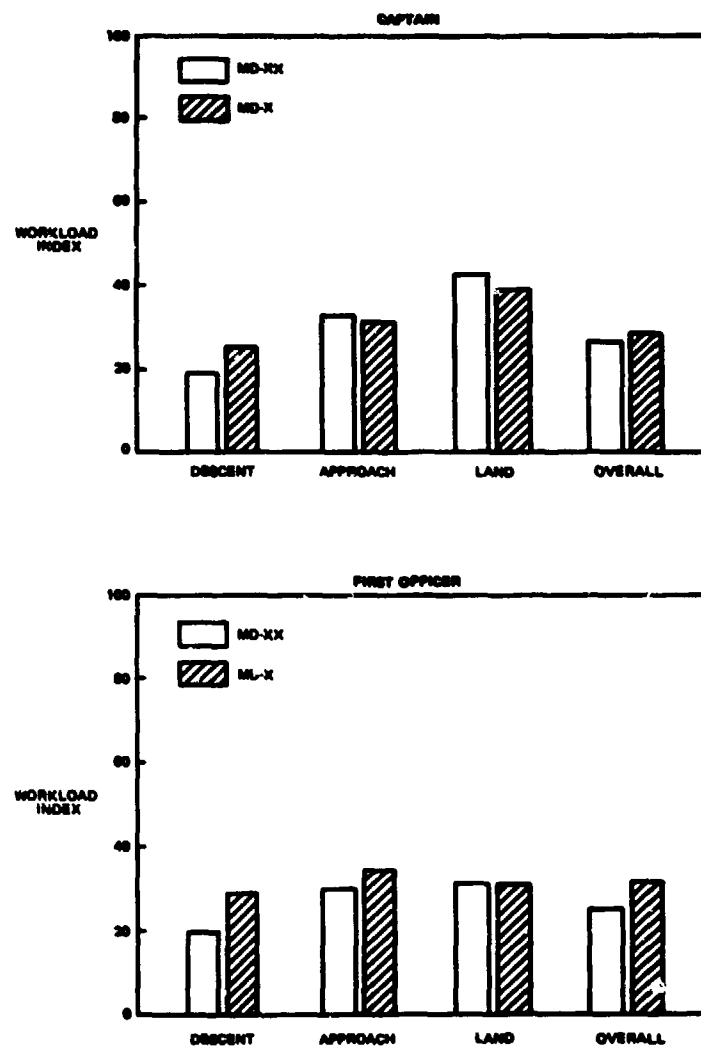


FIGURE 16. FLIGHT CREW WORKLOAD - DESCENT TO TOUCHDOWN

Officer's workload for the MD-XX is shown to be equal to or lower than on the MD-X, there appears to be some slight increase for the Captain. Further analysis indicated that the cause of this slight increase was as follows:

- 1 The MD-XX has an additional task, requiring the navigation display scales to be reset as the aircraft get close to touchdown.
- 2 During level-off, the altimeter in the MD-XX requires a slightly longer time to read and the flight data systems control display unit must be observed to cross-check the flight and navigation displays.

This analysis illustrates the manner in which the flight crew workload program can be effectively utilized. In this study, it was determined that the advanced configuration flight deck had slightly higher workloads during approach and landing than a conventional cockpit for the Captain's duties and an acceptable workload for the First Officer. The specific causes of the workload differential were subsequently established, allowing for redesign of equipment or a change in operational procedures to decrease the workload to acceptable levels.

The analysis does not stop at this point, however, but goes into more detail examining detailed flight segments and time breakdowns to ensure that, while average workloads are acceptable, there are no sharp peaks that are lost in the averaging. In addition, further study involves the imposition of contingency modes on the flight scenario to evaluate the workloads under these conditions.

Contingency Analysis

A contingency analysis is expressly designed to evaluate the impact of a degraded mode of operation on flight crew workload. This is accomplished by imposing an abnormal or emergency condition in each flight scenario used for the normal crew workload analysis and determining relative differences or changes.

For example, consider the situation in which one member of a two-member crew becomes incapacitated while in flight. Four steps must be taken to enable a safe landing:

- 1 maintain control of the aircraft;
- 2 take care of the incapacitated crew member
- 3 reorganize the flight deck; and
- 4 land the aircraft.

In this example, the First Officer becomes incapacitated during descent. The Captain's basic tasks remain unchanged, and he assumes as many of the First Officer's duties as is practical. The number of traffic advisories and communications with the Air Traffic Controllers (ATC) is the same as in the normal scenario. Additional verbal/aural tasks are inserted for communications with the ATC and company personnel to present the incapacitation as realistically as possible. Only those First Officer's tasks considered necessary for safety of flight are assumed by the Captain.

Two types of comparison are performed:

- 1 a new aircraft configuration with normal operating conditions versus a new aircraft configuration with degraded mode conditions; and
- 2 a new aircraft with degraded mode operating conditions versus a baseline aircraft with degraded mode conditions.

Examples of results by flight phase are shown in Figure 11. The new aircraft, the MD-XX, while having an increased workload for the Captain when his First Officer is incapacitated, does not overload the Captain. In the second comparison, when the new aircraft is compared to the baseline aircraft, the MD-X in the incapacitated crew member mode, a significantly lower workload is imposed on the Captain.

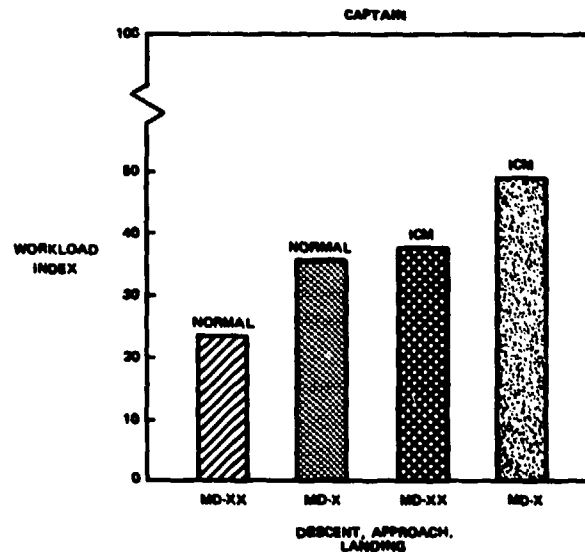


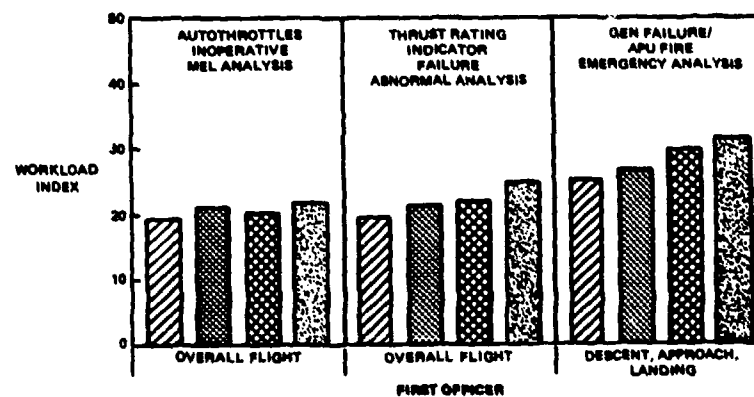
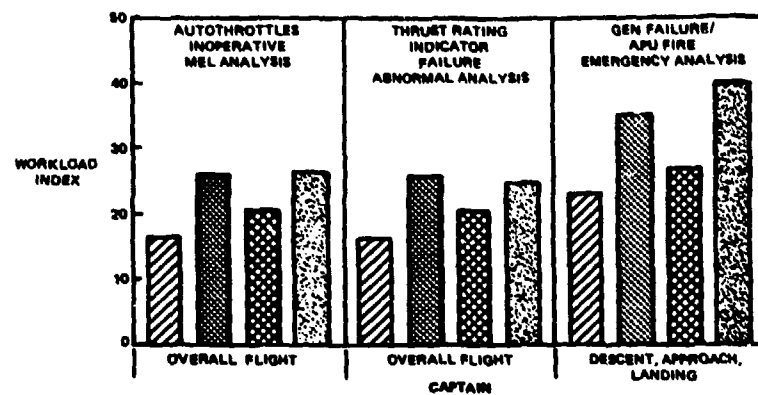
FIGURE 11. INCAPACITATED CREW MEMBER (ICM) WORKLOAD ANALYSIS

In addition, Figure 12 presents examples of the effect of other contingencies on average workloads during the flight. This indicates the versatility of the workload program and the variety of contingency situations which can be analyzed.

Subsystem or Equipment Analysis

Workload analysis may also be used as a design tool in the selection of a control and display layout for a particular subsystem. Figure 13 shows two proposed audio panel configurations for a modern jet transport. Audio Panel 1 represented the conventional panel with an on-off lever, and a separate control or volume adjustment.

In the second configuration, single continuous adjustment knobs incorporating push-on/push-off features are used for volume control. This pushbutton feature permits presetting the knobs to normal or to anticipated monitoring volume levels independent of the on-off function, a capability not available on Audio Panel 1. The time devoted to making volume adjustments may therefore be less with Audio Panel 2 than with Audio Panel 1.



MD-XX NORMAL MD-XX CONTINGENCY
 MD-X NORMAL MD-X CONTINGENCY

FIGURE 12. CONTINGENCY WORKLOAD ANALYSIS

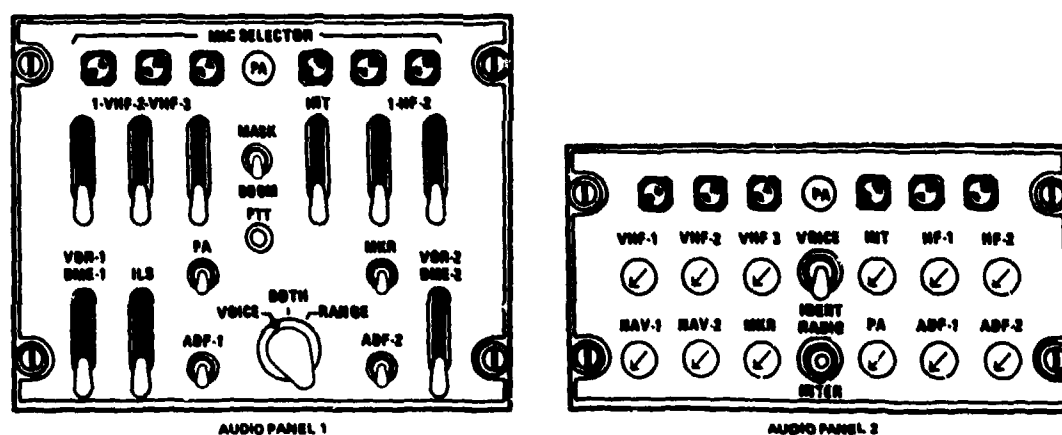


FIGURE 13. AUDIO PANEL WORKLOAD EVALUATION

TABLE 5
WORKLOAD RESULTS—
AUDIO PANEL EVALUATION

FLIGHT SEGMENT		WORKLOAD INDEX	
		CAPTAIN	FIRST OFFICER
TAKEOFF	AP1	1.98	8.88
	AP2	1.98	8.75
	Δ%	0.00	0.99
CLIMB	AP1	4.48	11.82
	AP2	4.48	10.71
	Δ%	0.00	-9.38
CRUISE	AP1	2.42	8.38
	AP2	2.38	4.87
	Δ%	-2.81	-12.68
DESCENT	AP1	13.24	27.31
	AP2	13.24	24.63
	Δ%	0.00	-10.11
APPROACH	AP1	8.29	9.71
	AP2	8.29	9.77
	Δ%	0.00	0.68
LANDING	AP1	0.00	0.07
	AP2	0.00	0.00
	Δ%	0.00	0.00
OVERALL	AP1	3.88	8.12
	AP2	3.85	7.38
	Δ%	-1.18	-9.05

This supposition is confirmed by examination of the numerical results of the workload evaluation presented in Table 5. In this case, Audio Panel 1 is considered the "standard" configuration and the results show reductions in the overall communications workload for the new system of approximately 1 percent for the Captain and 9 percent for the First Officer. Naturally, large workload reductions would be expected for the First Officer because one of his primary tasks is communications.

Another significant item extracted from this analysis is that workload reductions for the First Officer occur primarily during the climb and descent segments, which normally represent high workload phases of flight. Thus, any reduction in workload during these periods is especially beneficial. If the reductions occurred only during the low-workload cruise period and were of the low level shown for the Captain in Table 5, then the new development effort might be questioned.

Consequently, this comparative workload analysis of alternative audio control panel designs supports two conclusions:

1. the design for Configuration 2 shows superior workload characteristics over that of Configuration 1 and therefore is worthy of further development; and

in-flight communications workloads for future aircraft may be reduced by employing volume control designs which incorporate and on-off feature that acts independently of the volume level adjustment.

Certification Analysis

Flight crew workload analysis and design system can also be applied to aid in demonstrating compliance with Federal Aviation Regulations (FAR 25.1523) and its Appendix D (Minimum Flight Crew) (9). In this case, a comparative analysis is made between the new aircraft to be certified and an aircraft that has been operating in an airline environment for a number of years, is considered to have an acceptable level of workload, and has the crew complement certified under applicable Federal Aviation Regulations.

A study of this type is conducted to demonstrate how design differences in the crew station layouts, controls, and displays of the new aircraft affect flight crew workload during normal and degraded flight modes. The results for the normal workload are plotted in Figure 14. Overall reductions in workload are shown for the Captain and First Officer of the new aircraft equal to 32 and 7 percent, respectively. As indicated in Figure 14, there is a significant reduction in the captain's workload on the new aircraft in all flight phases, ranging from 26.8 percent during cruise to 44.6 percent during climb.

Additional analysis would be presented to the regulatory agency demonstrating the effect of abnormal and emergency flight situations on crew workload. An analysis of this type was submitted to the Federal Aviation Administration during the recent certification of the MD-80 aircraft.

Additional Analytic Capability

The task/timeline workload analysis methodology can also be applied as follows to all areas of aircraft development from the earliest concept through development, detailed design, certification, and crew training.

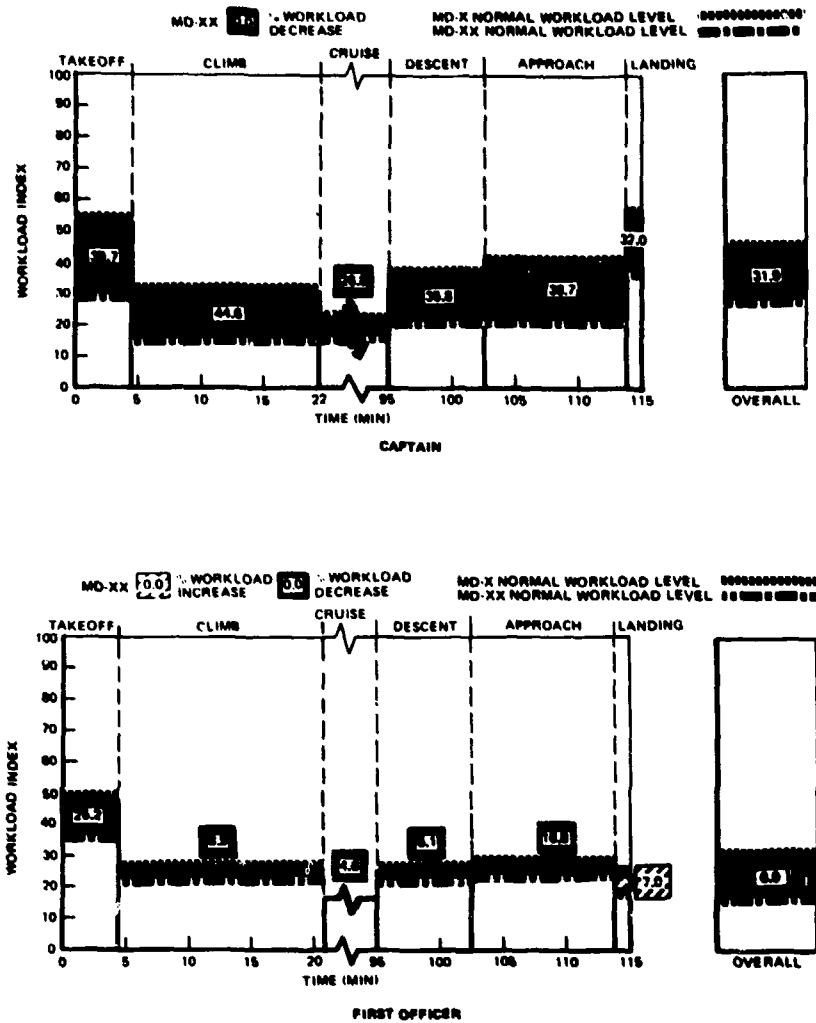


FIGURE 14. FLIGHT CREW WORKLOAD COMPARISON

- 1 Advanced design — As a tool in the creative stage of aircraft design to systematically determine such matters as allocation of functions to either a crew member or automation, and determination of the crew complement.
- 2 Design/development — For assistance in equipment placement, display format development, crew duty allocations, and operational procedures. During the design/development stage, the workload program may be used to design alternative design concepts in various trade studies involving different systems or subsystems.
- 3 Detailed design — The workload analysis process continues to verify crew duty allocation, the effects of contingencies on crew workload and mission completion success (or abort). Verification of the data base in the simulator mockup phase of development is also initiated. During this stage, when the design is frozen, the instructional development and training program is initiated, and the task listings, developed for the workload study, become useful in preparing training materials and flight manuals.

DISCUSSION

While there have been many symposia, papers and discussion groups devoted to the subject of workload, there seems to be no commonly accepted definition of the term. Because of this, there have been many different approaches to the qualitative and quantitative measurement of workload. The approach taken in this paper is concerned not so much with obtaining an absolute measure of workload — which would be highly desirable but is currently beyond our understanding — but with being able to use the comparative concept of workload measurement as a tool to aid in the design of work stations.

The task/timeline approach to workload quantification has certain limitations which preclude its being used in the true sense of a metric. In particular:

- 1 It does not consider cognitive or mental activities.
- 2 It does not take into account variations associated with ability and experience or dynamic, adaptive behaviour.
- 3 It cannot deal with simultaneous or continuous-tracking tasks.

At present, sufficient data do not exist on variations in task time associated with differences in operator capability or learning ability to include this factor in the analysis. Tasks are considered as being performed by an average operator.

With regard to simultaneous tasks, the workload program considers a serial approach to task performance and thus the results on this basis might be considered somewhat conservative. Continuous-tracking tasks are handled by an assumption of serial task performance. For aircraft control wheel or throttle continuous-input tasks, flight test data were examined to determine pilot discrete inputs to these controls. Averages from these data on frequency and duration may then be used in the analysis.

Admittedly, all of these compromises do not allow for the expression of an absolute metric of workload. In fact, there is no universal agreement in the industry as to what levels, derived, from task/timeline analysis, are considered acceptable — whether the level be overload or underload.

No accepted method has been developed to adequately compensate for these limitations. Subjective assessment or simulator studies are sometimes used to help improve insights into the significance of these factors. In general, we support this approach to improving the understanding of human ability in system operation. Each approach has its value. To use one is not to deny the value of the other.

The task/timeline approach to workload analysis which is described in this paper, however, was subject to close scrutiny by many agencies because of the controversy over a two-member flight crew. The following comment from a presidential task force is considered significant (10).

"At present, the only generally accepted method for evaluating workload is task/timeline analysis based on comparison with previous aircraft designs. This technique, supplemented by improved subjective evaluation methods applied by qualified pilots, will offer the best means for demonstrating compliance with FAA crew complement criteria."

The comparative concept provides a basis for extensive use of this methodology and, in fact, allows for a wide range of evaluation of variations in work station design. Comparisons can be made between different aircraft, systems, or individual pieces of equipment, or even to examine the effectiveness of different panel locations for controls or displays.

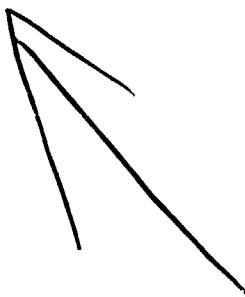
If the baseline used in the comparison is considered to have an acceptable workload, then the analysis will indicate which has the lowest workload and by what magnitude. Even when used in a noncomparative mode, the technique allows for the assessment of those portions of scenario where workload levels can be expected to be substantially higher than the average, and thus allows for more detailed analysis aimed at minimizing peak workloads. Another plus is the fact that the procedure can be applied early in the design cycle and thus have the ability to influence design. Though mockups and simulators would be advantageous in establishing crew procedures, they are not absolutely required in the analytical process.

A typical workload analysis on a new aircraft or work station is considerably labor-intensive in that extensive task listings describing detailed operation of the system under consideration must be prepared. Moreover, a number of different scenarios or missions may have to be considered. Once the baseline is developed, however, it can then be modified to reflect various concepts or design options with little difficulty. It is fairly evident, however, that the only way to accomplish an analysis of this magnitude is with an automated facility. Machine computational capabilities plus the flexibility of the technique allows for extensive graphic presentation and facilitates analysis.

An effort is currently underway to improve the computer program and its input software. The new program will automatically generate various scenarios by supplying formatted flight segments with their associated time factors, and provide simplified input formats for task generation. It will contain an extensive library of system procedures which will allow for rapid computation of task time. In addition, consideration is being given to adapting methodology developed for the assessment of human reliability for the program, thus providing an additional measure of human performance to supplement the workload analysis.

REFERENCES

- 1 CHILES W D Objective methods for developing indices of pilot workload. Report No FAA-AM-15, FAA Department of Transportation, Washington DC, July 1977
- 2 GREENING C P Analysis of crew/cockpit models for advanced aircraft. Report No NWC TP 6020, Naval Weapons Center, China Lake, California, February 1978
- 3 CHILES W D On the specification of operator or occupational workload with performance-measurement methods. Human Factors, Vol 21, No 5, October 1977
- 4 WILLIGES R C Behavioural measures of aircrew mental workload. Report No MDCJ5291, Douglas Aircraft Company, September 1971
- 5 WIEWILLE W W Task analysis methods: Review and development of techniques for analyzing mental workload in multiple task situations. Report No MDCJ5291, Douglas Aircraft Company, September 1971

- 6 BROWN E L
STONE G
PEARCE W E
Improving cockpits through crew workload measurement. Report No MDC 63-55, Douglas Aircraft Company, 1975
 - 7 MUNGER S J
SMITH R W
PAYNE D
An index of electronic equipment operability — data store. Report No AIR-C43-1/62, American Institute for Research, Pittsburgh, Pennsylvania, December 1962
 - 8 STONE G
REGIS E R
GULICK R K
Data base validation, DC-9 Super 80/DC-50. Cooperative flight crew workload study. Report No MDC J8748, Douglas Aircraft Company, June 1980
 - 9 ANON
Federal Aviation Regulations, Part 25, Airworthiness Standards: Transport Category Airplanes
 - 10 McLUCAS J L
DRINKWATER F J
LEAF H W
Report of the President's Task Force on Aircraft Crew Complement. Prepared for the President, The White House, Washington DC, July 1981
- 

CHAPTER 4

PILOT SUBJECTIVE EVALUATION OF WORKLOAD DURING A FLIGHT TEST CERTIFICATION PROGRAMME

by

Frank T Ruggiero
and
Delmar M Fadden
The Boeing Company
Seattle Washington 98124-2207
USA

INTRODUCTION

To date there is no agreed upon definition of mental workload and therefore there is no agreement on how it should be measured (1). Current workload researchers do seem to agree on at least three aspects of mental workload: it is a multidimensional construct, a clear distinction must be maintained between imposed mental load (task load) and the mental load as experienced (subjective load), and the use of subjective ratings should be central to any investigation of workload (2). On this last point, The President's Task Force on Aircraft Crew Complement made the following recommendation: "This technique (task/timeline analysis based on comparison with previous aircraft designs), supplemented by improved subjective evaluation methods applied by qualified pilots, will offer the best means for demonstrating compliance with FAA crew complement criteria. We recommend that FAA incorporate such methods in the tests to be employed for the certification of the B-757 and B-767 aircraft", (3).

The paper outlines the Pilot Subjective Evaluation (PSE) process developed by Boeing, in conjunction with the FAA, to supplement the analytical, simulator, and flight test crew workload evaluation techniques used to demonstrate compliance with the minimum crew size requirements of FAR 25.1523 and Appendix D (4).

767 WORKLOAD EVALUATION

The workload assessment techniques used in the design, development, and certification of the Boeing Model 767 airplane addressed two basic issues: timeliness of crew actions and ease of operation. To be acceptable for certification both the nature and timing of crew tasks must be well within the range of demonstrated pilot capacity and the sequencing of tasks must allow sufficient reserve time to accommodate unexpected events. A three part process was used to ensure that the final design would satisfy these requirements. Analysis provided an early indication of the suitability of "paper" designs. Part-task simulation provided a detailed look at specific man-machine interactions and, when the design has progressed far enough, verification of the operational suitability of the integrated design. The final check was a flight test demonstration in the actual operational environment.

Analytic techniques are of particular value to the aircraft manufacturer since they offer the potential for identifying and correcting workload problems early in the design phase, when the cost of change is relatively low. The analytic techniques which we have found to be most useful focus on traditional time and motion evaluation. These techniques give preliminary indications of task loading and timing. They also permit comparative evaluations of panel layouts and operating procedures. The results are characterized in terms of the time required to accomplish the various hand and eye tasks associated with operating the airplane. A portion of the mental effort associated with these tasks is addressed through an information theoretic technique (5) which quantifies the information exchange between the pilot and the airplane operating environment.

High fidelity simulations of the aircraft and flight deck permit both objective and subjective evaluations of the workload associated with new concepts and design features. The initial 767 airplane simulation activities concentrated on specific features of the primary flight displays and the flight management system. Later in the program more complete representations of the flight deck were used to determine the effect of the electronic displays on the pilot's scan pattern during routine manual flight operations. The results showed that neither instrument dwell time nor scanning strategy were likely to be significantly altered by the displays and display formats planned for the 767. Simulation was also used as a link between analysis and flight test providing data which made possible correlation of the objective analysis results with the largely subjective flight test results.

Flight testing of the flight deck is done to validate the earlier analysis and simulation results and to check the effect of subtle factors in the operational environment which cannot be duplicated on the ground. Initial flight testing was primarily developmental... a means to finalize certain design characteristics and to document the airplane performance. Later testing was aimed specifically at showing compliance with the applicable FAA regulations.

Since flight testing is expensive and time consuming, every effort was made to integrate tests. A limited amount of flight testing was conducted with a fully instrumented airplane and a video cockpit monitoring system to provide data for comparison with simulation results. However, primary emphasis was placed on pilot assessments using a nonintrusive questionnaire process as the measurement instrument.

PILOT SUBJECTIVE EVALUATION

Subjective assessments have been used as part of the evaluation of all modern transport aircraft. In the case of the Boeing 737 and the McDonnell Douglas DC-9-80 these assessments were formalized as part of the certification record. The President's Task Force recommendation of "an improved subjective evaluation" was the subject of considerable discussion

AD-P905631

within Boeing and with the FAA. The criteria established for development of this improved evaluation were: (1) results must relate to the workload functions and factors of FAR 25.1523 and Appendix D, (2) pertinent conditions for the phase of flight must be identified, (3) the results should complement the other comparative workload assessment techniques in use, and (4) the evaluation methodology must be compatible with the realities of a major flight test program. These criteria were best satisfied by development of a questionnaire process to be completed by the Boeing and FAA flight test evaluation pilots.

PSE Development

The development steps for the Pilot Subjective Evaluation (PSE) involved numerous cycles of question design, round-table discussions with a cross section of pilots to determine suitability, and simulator/flight test trials to refine the wording and format. Finally, a validation study involving Boeing and FAA pilots was conducted using the PSE on 166 flights mid-way through the 767 flight test program. Only then was the PSE considered adequate for use during the Minimum Crew Size flight tests.

The PSE questionnaire covers the departure phase, from takeoff to cruise, and the arrival phase, from the beginning of descent through landing. Any nonnormal procedures encountered were also evaluated regardless of flight phase. The flight phase dependent questions began by establishing certain facts about the flight conditions, interactions with ATC, and flight equipment usage. This was followed by questions asking the pilot to compare specific aspects of workload flight functions on the 767 airplane with similar activities on a reference airplane. The flight functions were those specified in FAR 25 Appendix D plus Flight Management System Operation and Monitoring. The choice of reference airplane was left up to the pilot; however, his choice was identified on the questionnaire. The PSE process was completed with a debriefing interview which solicited pilot comments about each flight to clarify subjective ratings.

The workload characteristics initially associated with each flight function were: mental effort, physical difficulty, and time required. During the PSE development phase it became evident that certain combinations of these characteristics and specific workload flight functions were not possible for the pilot to evaluate and they were not measured. For example, the tasks of engine/airplane system operating and monitoring, along with manual flight path control and communications involve actions which are so highly distributed that most pilots felt time estimates would be meaningless even on a comparative basis.

For the workload flight functions of command decisions and collision avoidance the Boeing and FAA pilots suggested that time available would be a better measure than "time required". Physical difficulty was readily understandable with all of the workload flight functions except command decisions and collision avoidance. Similarly, mental effort was readily associated with all but two of the workload functions.

All pilots who participated in PSE development felt that neither communications nor collision avoidance should include a mental effort rating. The largest changes in the 767 flight deck, when compared to previous transport aircraft, are those designed to aid the pilot with navigation and command decision making. It was decided to add a subjective rating for the effectiveness of these changes to complete the assessment of normal operations.

PSE Administration

Each rating was made by the pilot marking the appropriate box on a seven point adjective scale. The middle box, marked "same", represented a workload equivalent to that experienced on the reference airplane when operated in similar circumstances. The three boxes on either side represent progressively more or less workload than on the reference airplane. The adverbs slightly, moderately, and much were chosen because of their measured equal spread of meaning (6) and used to identify the boxes on both the less and more workload sides. It was decided to orient the scales with the "better than" boxes to the right in all cases. Since the questionnaire was used frequently by each of the evaluation pilots, there was no advantage to an alternated scale orientation. This orientation of the scales resulted in less errors by pilots and was easier for interviewers to scan pilot responses during the debriefing interview.

The questionnaires were completed immediately after each flight segment. Pilots took as much time as needed to complete the form, without interruption. Debriefing interviews were conducted at the end of the sequence of flights for the day. The interviewer asked about any "worse-than-reference" airplane ratings and any differences between departure and arrival ratings. The interviewer also recorded any other pilot comments about the ratings, the PSE process, or the airplane in general. Completed questionnaires and comment sheets were then coded for data processing and analysis.

PSE and minimum crew size flights

Dedicated minimum crew size flights involved 7 different pilots, 10 different airports, 32 daytime and 18 nighttime operations, and approximately 40 flight hours. During 80% of these flights, inoperative or failed equipment was intentionally introduced. The 737 airplane was listed as the reference airplane for six of the pilots, one pilot listed the 727 as his reference airplane. All seven pilots answered "yes" to the question, "Have you flown your reference airplane (or an approved simulator for that airplane) in the last 90 days?"

A summary of the flight condition data is shown in Table 1. The information collected for Departures and Arrivals has been combined. The categories of ATC interaction and Flight Modes used show a wide range indicative of operating the airplane in a manner representative of scheduled airline operation.

1. Flight Conditions			
a. Airport Weather (Ceiling and Visibility)	b. Precipitation at Airport	c. Meteorological Conditions Aloft	d. Turbulence
2 less than 400ft and 1 mile 16 400ft and 1 mile to 1000ft and 3 miles 82 Better than 1000ft and 3 miles 100%	94 None 0 Light 1 Moderate 0 Heavy 100%	65 VMC 9 IMC 26 Mixed 100%	66 None 28 Light 6 Moderate 0 Severe 100%
2. ATC Data			
a. ATC Procedures	b. Did you enter an Amended Route into the FMC/CNU	c. Level of interaction with ATC	d. Number of Altitude clearance changes
10 VFR Only 34 IFR: Vectoring Only 10 IFR: Assigned Route 46 IFR: Vectoring + Assigned Route 100%	67 Yes 33 No 100%	44 Low 53 Moderate 3 High 100%	22 1-2 67 3-4 7 5 or more 4 None 100%
3. FMS Modes Used			
a. EHSI Use	b. Autopilot Use	c. Flight Director Use	d. A/T Use
79 Map 7 VOR/ILS 12 Both 2 Neither 100%	72 CMD 0 CWS 18 Not used 100%	22 Full-Time 33 Part-Time 45 Not used 100%	47 Full-Time 28 Part-Time 25 Not used 100%

TABLE 1. RELATIVE PERCENTAGE FLIGHT CHARACTERISTICS COMBINED FOR DEPARTURES AND ARRIVALS OF B-767 AIRPLANE MINIMUM CREW SIZE FLIGHT TESTS

The absolute frequencies for the pilots' subjective ratings of the workload associated with Normal Operations: Departure and Arrival are combined and shown in Table 2; there were no reliable differences between Departure and Arrival ratings. The rating category which contains the 50th percentile rating (median rating) is marked with an arrow. In all cases, the median rating for the 767 airplane was equal to or better than that for the reference airplane.

There were 1.3% of the items marked to the left of "same as reference airplane" and these ratings were all in the "slightly" more workload category. Pilots were instructed to complete only those items which were applicable to their flight duties. The NA (not applicable) frequencies represent situations where one of the crew members did not have flight duties associated with that workload function for a given Departure or Arrival.

The subjective questionnaire results correlate well with earlier analysis and simulation results. Pilot comments also substantiated the pattern of results shown in summary Table 2. For example, hand and eye motion data from the timeline analysis predicts that the 767 will exhibit slightly lower physical workload than the 737 for the primary pilot tasks involving actuation of controls.

Many researchers claim a strong correlation between mental workload and time stress. (7) The workload analysis package includes two measures related to time stress. The timeline provides estimates of workload in terms of the ratio of time required to time available. The task-time probability measure identifies time critical or overlapping tasks. These analytic results compare well with the pilots' assessments of mental workload. Further validation of the PSE process was obtained from the 166 preliminary 767 airplane flights where changes to various equipment were reflected in concomitant changes to the workload rating for related functions.

CONCLUSIONS

The multidimensional nature of workload, as it is currently understood, does not lend itself to a simple numerical analysis yielding a single index of workload. In fact, for most design related applications, multidimensional results provide a better picture of the real world situation and give some indication of what area may need improvement.

The total package of workload assessment techniques applied to the 767 program was successful in supporting the design, development, and certification process. The addition of formalized subjective measures to the traditional objective analyses provided information validating the analytic and simulation based estimates of physical workload and complementing the estimates of mental workload.

In keeping with current constructs of pilot workload, the Pilot Subjective Evaluation is multidimensional. The PSE avoids the vagaries of absolute workload assessments by asking the pilot to compare specific aspects of workload on the subject airplane with a familiar, previously certified, reference airplane. This technique is nonintrusive and readily applicable to a flight test operational environment.

Development of the PSE and the procedures for its use have been subjected to intense scrutiny including extensive flight test. While other aircraft designs may necessitate alteration of some of the workload task functions, the basic approach should provide useful in many applications. As we progressed through the development of this technique, we were surprised at the lack of documented experience with multidimensional, comparative, subjective measurement. We hope that our success with the technique will encourage others to attempt additional applications and that more basic research in this field will yield more powerful methods for the analysis of the results.

REFERENCES

- 1 HART S Defining the subjective experience of workload, Proceedings of the Human Factors Society, pp 527-531, 1981
- 2 EGGEMEIER F T Current issues in subjective assessment of workload Proceedings of the Human Factors Society, pp 527-531, 1981
- 3 McLUCAS J L
DRINKWATER F J
LEAF H W Report of the President's Task Force of Aircraft Crew-complement p. 8 July 1981
- 4 ANON Code of Federal Regulations, Title 14: Parts 1-59, pp 367 and 386, January 1982.
- 5 SENDERS J W The human operator as a monitor and controller of multidegree of freedom systems, TIEE Transactions: on Human Factors in Electronics, pp2-5, 1964
- 6 PACS B M
SCIO W F
NNER E J Magnitude Estimations of expressions of frequency and amount. J App Psychol 59(13) pp 13-320
- 7 M LAY N Subjective mental workload. Human Factors, Vol 24(1), pp25-40, 1982

Normal Operations

	Mental Effort	Physical Difficulty	Time Required	Understanding of Horizontal Position
Navigation	More Less 0 0 5 21 27 35 10 N.A. = 2	More Less 0 0 1 40 26 22 9 N.A. = 2	More Less 0 0 3 33 23 30 9 N.A. = 3	More Less 0 0 3 16 9 4 2 27 N.A. = 3
FMS Operation and Monitoring	More Less 0 0 2 22 27 13 9 N.A. = 27	More Less 0 0 0 24 32 8 9 N.A. = 27	More Less 0 0 2 22 29 10 10 N.A. = 27	Blank
Engine/Airplane Systems Operating and Monitoring	More Less 0 0 0 27 44 16 11 N.A. = 2	More Less 0 0 0 36 38 13 11 N.A. = 2	Blank	Blank
Manual Flight Path Control	More Less 0 0 1 49 11 1 7 N.A. = 30	More Less 0 0 1 55 4 3 7 N.A. = 30	More Less 0 0 0 38 33 3 9 N.A. = 17	More Less 0 0 0 26 17 2 13 N.A. = 17
Communications	Blank	Blank	Blank	Blank
Commanded Decisions	More Less 0 0 0 42 27 7 9 N.A. = 18	Blank	More Less 0 0 0 55 29 4 9 N.A. = 3	Blank
Collision Avoidance	Blank	Blank	Blank	Blank

(Item 4.3 to be completed for EICAS equipped airplanes only.)

(Item 4.4 to be completed for manual flight only.)

(Complete these items for 4.6 and 4.7 only.)

Usefulness of Information

CHAPTER 5

THE USE OF SUBJECTIVE WORKLOAD ASSESSMENT TECHNIQUE IN A COMPLEX FLIGHT TASK

by

F V Schick
DFVLR Institute for Flight Guidance
Flughafen, D-3300 Braunschweig
Federal Republic of Germany

and

R L Haan
US Air Force Aerospace Medical Research Laboratory (HEC)
Wright Patterson AFB
Dayton, Ohio 45433, USA

INTRODUCTION

With the increasing tendency towards all-digital airborne and groundbased workspaces, the search for satisfactory mental workload measurement methods has become one of the most active human factors research areas. Designers and engineers have asked for better methods to assess mental workload at all stages of system development — but especially in the high-fidelity simulator and in actual in-flight tests.

Techniques for measuring mental workload (hereafter referred to merely as "workload") can be divided into three basic categories:

- 1 physiological,
- 2 behavioural, and
- 3 subjective.

The present paper deals with one particular technique belonging to the third group of methods, which always use some form of operator self-report (eg rating scales or questionnaires). The subjective methods seem at first glance to be almost too simple and "unscientific". However, as Johannsen (1) has noted, if an operator *feels* his workload level is high then it is high, regardless of what other measures show. Indeed, it may be the only meaningful definition of mental workload, he says.

Some of the criteria normally applied in evaluating the various workload techniques are: non-intrusiveness, ease of implementation, operator acceptance, and sensitivity to variations in task demand. Although the subjective techniques tend to satisfy these requirements, probably better than behavioural and physiological methods, they have exhibited a couple of undesirable characteristics. First of all, in most applications of the technique the scales used are specific to a single investigation and therefore not validated for general use. Secondly, there is little evidence that workload rating scales have been developed on the basis of psychometric theory, eg, Williges and Wierwille (2). The result is that most available scales have unknown metric properties and, at best, provide only ordinal measurement capability. Because of this, the variety and power of available statistical analyses are limited.

In order to deal with these undesirable properties of subjective methods, a procedure known as the Subjective Workload Assessment Technique, or SWAT, was developed at AFAMRL by Reid and his colleagues (3) (4) (5). In SWAT, subjective workload is defined as being composed of three dimensions;

- 1 time load,
- 2 mental effort load, and
- 3 psychological stress load;

they are an adaptation of those suggested by Sheridan and Simpson (6). (It is generally agreed among researchers today that no single dimension is capable of describing workload.) Each dimension is represented by an individual three-point rating scale with a description for each level of load. SWAT is based on conjoint measurements and scaling (eg Krantz and Tversky (7)) and permits ratings on three dimensions to be combined into one overall scale of workload. In order to identify the appropriate mathematical rule for combining the three dimensions into one overall scale, a "scale development" phase is completed. During this phase, subjects rank order the subjective workload associated with the 27 possible combinations that result from the three levels of time, mental effort, and stress load. After completion of scale development, an "event scoring" phase is started. This phase is the actual experiment, during which the subjects perform the task(s) of interest and rate the time, mental effort, and stress load imposed by performing the task. This three-part rating corresponds to one of the 27 workload statement combinations from the scale development phase — and therefore to one of the derived interval values on the SWAT scale. The SWAT value for that time, mental effort, and stress load combination become the datum for inclusion in the usual statistical analyses.

An important part of developing any new measurement technique is scale validation. Since workload is a hypothetical construct, that is, not a directly observable phenomenon, validation is accomplished by comparing a new measure with other measures already in use.

AD-P005 632

The following description concentrates on the application of SWAT in a validation experiment, which consisted of a series of landing approaches flown in a moving cockpit simulator. The relations of SWAT to other measurements taken in the experiment will be published in a separate report. The SWAT version used was a recently developed German-language version. However, since a previous investigation provided strong support for the assumption that the German version was an accurate equivalent of the original, the results should hold for SWAT in general.

2 METHOD

In order to provide a suitable task and environment, a simulated flight task was selected. The task consisted of a ten minute-flight in Hamburg terminal area, including approach and landing; it was flown manually in a moving cockpit simulator. The difficulty of the piloting task was changing along the prescribed flight path. Various combinations of task elements, which were, eg, straight and level flight, curve, descent, deceleration, and ILS intercept, made a subdivision of the entire flight profile into six segments with different levels of task demand (or workload, respectively) possible. The flight task was flown 14 times by each subject, prior training runs not included. On one half of the 14 runs, a simulation of a strong low-altitude wind shear became active in the final flight segment. This was used to present an additional variable of task difficulty to the pilot, with its own two clearly distinct levels of workload. To assure that the windshear, used alternatively with the no windshear condition in the final segment, was actually a significant load factor, one item of an eleven-item questionnaire which was filled out by each subject at the end of his experimental session, required the pilot to give an estimate of the difference in task difficulty between these two conditions. Another item of the questionnaire required the pilot to put the first five flight segments in a rank order of task difficulty.

2.1 Subjects

A total of fourteen pilots were selected for participation in this validation study. They all had commercial pilot licences and were licensed for flights following instrument flight rules (IFR). Most of them were test pilots. Their flying experience ranged from 800 to 9500 hours.

2.2 Flight Scenario and SWAT Rating Procedure

The approach and landing at the airport of Hamburg had to be flown manually, without the assistance of a copilot. The simulation runs started 10 miles from the initial approach fix, Hamburg VOR. On a randomly alternating basis, the pilot had to take over control of the airplane either 10 miles south of the Hamburg VOR, with inbound course 009, or 10 miles west of the fix, with inbound course 069 (see Figure 1). He was partially guided by "radar vectoring" to the ILS runway 23. The ATC scenario is given in Table 1.

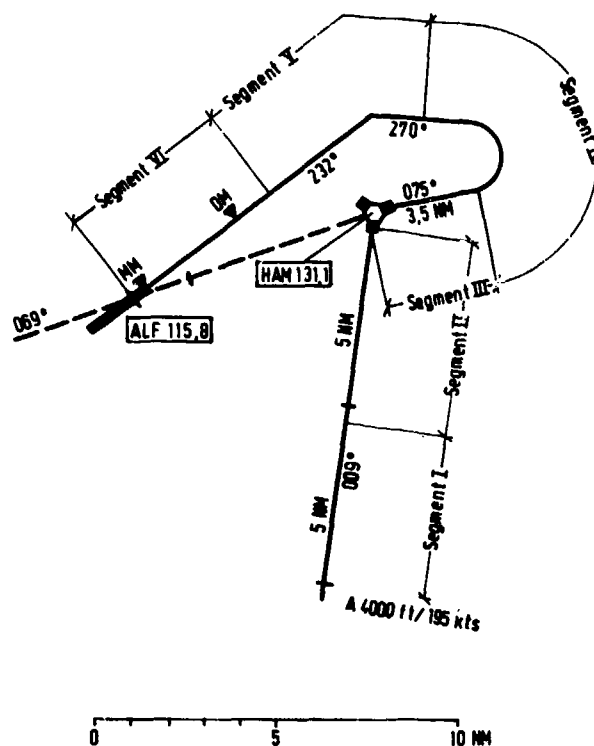


Fig.1 Flight task geometry

Table 1 Scenario for SWAT-Experiment

Fix/Altitude	Heading	Elision Time	Segment	Description of Activity	R/T-Communication (minimum of 3 sec before activity starts) Radar-vectoring L: Local; P: Pilot
10NM to MAN (113.1), A040	009 or 0 069	0	I	Speed 210Kts, holding straight and level	
7.3NM to MAN		0.75			
6NM to MAN		1.15		SWAT Rating (1)	L: DPL-A descend to altitude 3000ft, speed 180 kts ... and your SWAT ratings please P: DPL-A leaving 4000 for 3000, speed 180 kts ... by SWAT rating is ...
3NM to MAN		1.43	II	Starting to descend	
2.3NM to A040 MAN				Descending	
1NM to MAN A034		2.66		SWAT Rating (2)	L: DPL-A turn right to intercept radial 075 now ... and your SWAT rating please P: DPL-A turning right ... my SWAT rating is ...
.5 1.5					
Over MAN A030	075	2.97	III	Turning from heading 009 or 069 to head- ing 075. Maintaining 3000 ft.	
1NM from MAN		3.3		Holding straight and level	
3NM from MAN A030		3.96		SWAT Rating (3)	L: DPL-A turn left heading 270, descend to altitude 2200 ft ... and your SWAT rating please P: DPL-A DPL-A turning to 270, altitude 2200 ... my SWAT rating is ...
3.5NM from MAN A030		4.13	IV	Starting to turn and starting descent	
					L: DPL-A cleared ILS 23 P: DPL-A cleared ILS 23
Crossing MAN VOR rad- ial 045, A022	270	5.03		Crossing MAN VOR radial 045, heading 270, level 2200 ft	
Crossing MAN VOR rad- ial 025	270	5.34		SWAT rating (4)	L: DPL-A report localizer established ... and your SWAT rating please P: DPL-A wilco ... my SWAT rating is ...
Crossing MAN VOR rad- ial 010		5.86	V		
ON radial 232; ILS DLM 110.1				Intercepting ILS	P: DPL-A localizer established L: DPL-A roger, contact tower 126.85 P: DPL-A tower 126.85
		6.15			P: Hamburg tower, this is DPL-A L: DPL-A report outer marker P: DPL-A wilco
7.2M from ILS DLM		7.08		SWAT rating (5)	L: Your SWAT rating please P: My SWAT rating is ...
7NM minus 0.5 NM from ILS DLM		7.29	VI	Starting to descend on the glide slope	
					P: DPL-A passing out marker L: DPL-A roger, cleared to land P: DPL-A cleared to land
				Widesehear portion of flight	
0.5NM from ILS DLM		9.86		SWAT rating (6)	
Touchdown DLM		10.07			L: Your SWAT rating please P: My SWAT rating is ...

The flight was divided into 6 segments. Each segment was defined by specific partial tasks. In the beginning of segment 1 the pilot took over the "clean" plane at 4000 ft and with an airspeed of 195 kts. The segment ended with the instruction to descend to 3000 ft, to reduce speed to 180kts and to give the SWAT-rating for the first segment. About 1 mile before the Hamburg VOR the second segment ended with the instruction to turn right and to give the second SWAT-rating. After intercepting radial 075 and about 3 miles away from the VCR, the pilot was instructed to turn left to heading 270, to descend to 2200 ft and to judge segment 3 with a SWAT-rating. When heading 270 was established and the airplane had reached the altitude of 2200 ft, the pilot gave his SWAT-rating for segment 4 and was cleared for ILS-approach on runway 23. About two miles to the Outer Marker the pilot gave the SWAT-rating for segment 5, which essentially consisted of the interception of localizer and glidepath. In half of the flights, after passing the Outer Marker, the pilot had to deal with a windshear situation. This last segment 6 had to be judged with a SWAT rating immediately after touchdown.

During the flight, the pilot had to do all R/T communication himself. He also had to take care of the flap setting and the landing gear. The subjective workload estimates were taken at the end of each flight segment. The experimenter, who also did the ATC communication with the pilot, always asked him to give his workload rating for the segment just completed. The pilot did so by making a verbal call-out of a three-digit combination, where the first digit always referred to time load, the second digit to mental effort load, and the third one to psychological stress load. As a reminder of the required sequence of these three dimensions of SWAT ratings, a small placard was fixed on the lower instrument panel. The SWAT rating call-outs, as well as all other communications, were recorded on audio tape. Additionally, the experimenter wrote down the SWAT ratings in his "test log".

3 RESULTS AND DISCUSSION

To make sure that the segmenting of the flight task, as it was done here, actually allowed the desired distinction between varying levels of load, the difficulty estimates given by the pilots in the post-experiment questionnaire were checked prior to the analyses of the SWAT data.

First, as it was expected, windshear in the final segment made the piloting task more difficult, at least "somewhat more" (three quotations), but more frequently "much more difficult" (eleven quotations). Second, the rank orders given for the task difficulties of the first five flight segments indicated a stepwise, monotonic increase of workload from segment 1 to segment 5 (see Figure 2 for mean ranks). There was a high agreement between the individual rank orders of the 14 pilots. This is reflected by a high coefficient of concordance $W = .846$, which is also significant beyond the one per cent level (chi square 0 47,37 at $df = 4$). Taken altogether, it indicates that the subtask segments defined here represented clearly distinct levels of the independent variable, which should then be reflected by the SWAT technique, too.

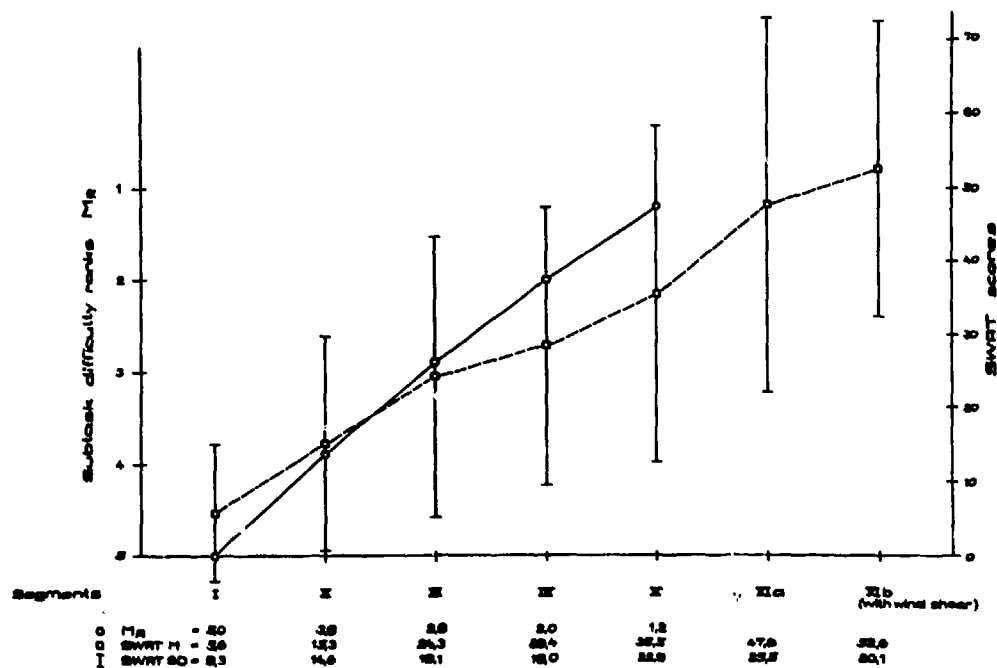


Fig.2 Mean of difficulty ranks M_R (solid line), SWAT score means M (dashed line) and Standard Deviations SD (vertical bars) obtained from 14 subjects in the various segments of the flight task

As an initial step for the analysis of the SWAT data, the transformation of the three-dimensional ratings into one overall scale was performed. Under the assumption that the calculated overall SWAT scale is valid for all pilots involved in this experiment, two-way analyses of variance (with pilot as factor one and segments as factor two) were carried out. The first Anova, using only the data obtained in the final segment, under windshear/no windshear conditions, was done to check the validity of SWAT to discriminate between two quite different workload levels of basically the same task. The second Anova, using the data obtained in the first five segments, analysed the SWAT measurement's ability to discriminate between more than two different subtasks, which may have much less apparent workload differences between them.

As a finding from both analyses, SWAT ratings showed highly significant differences between the individual pilots. The large standard deviations of the SWAT scores (see figure 2) appear to be mainly ascribable to this high interindividual variation.

But, nevertheless, in the Anova carried out first, the discrimination of the final approach segment six, ie wind shear and turbulence influenced flight vs. no disturbances, was also significant (probability of $F = 6.01$ (df:1; 140) was less than five per cent).

Additionally, the following Anova showed that the SWAT score differences between the first five flight segments were also significant (probability of $F = 179.9$ (df:4; 770) was less than five per cent, too). The data showed a monotonic increase of the SWAT ratings from segments one to five, which corresponds well with the task difficulty levels administered in the experiment. Moreover, Scheffe post-hoc comparisons of means showed that *all* pairwise differences between any two successive segment SWAT scores were also significant beyond the five per cent level.

So, it can be concluded from this experiment that basic evidence was found for the validity of SWAT as a tool for the assessment of mental workload, and thus for its applicability in high-fidelity simulation and actual in-flight tests.

It should be mentioned, however, that the collection of SWAT data has to be planned carefully, in order to ensure that calls to do SWAT ratings do not interfere with the flight task (Pilot comments indicated that, eg, giving SWAT ratings immediately after receiving ATC instructions might be viewed by some subjects as being a distraction from the flight task. From this point of view, the collection of SWAT ratings at fixed, rigid time intervals appears to be not a good practice. Rather, the SWAT data collection should be organized as being event-related, in a way which assures that both, subject and experimenter, have the same understanding as to which task or sequence of tasks just accomplished a SWAT rating refers to.

REFERENCES

- 1 JOHANNSEN G et al Final report of the experimental psychology group. In N Moray (Ed) Mental Workload: Its theory and measurement. New York: Plenum Press, 1979
- 2 WILLIGES R WIERWILLE W Behavioural measures of aircrew mental workload. Human Factors, 21, 549-574, 1979
- 3 REID G SHINGELDECKER C EGGEMEIER F Application of conjoint measurement to workload scale development. Proceedings of the 1981 Human Factors Society Annual Meeting, 522-526, October 1981
- 4 REID G SINGELDECKER C NYGREN T EGGEMEIER F Development of multidimensional subjective measures of workload. Proceedings of the 1981 IEEE International Conference of Cybernetics and Society, 403-406, 1981
- 5 REID G EGGEMEIER F SHINGELDECKER C Subjective workload assessment technique. Paper presented at the 1982 AIAA Workshop on Flight Testing to Identify Pilot Workload and Pilot Dynamics, January 1982
- 6 SHERIDAN T SIMPSON R Toward the definition and measurement of mental workload of transport pilots. Cambridge, Massachusetts: Massachusetts Institute of Technology Flight Transportation Laboratory Report R 71-4, January 1979
- 7 KRANTZ D TVERSKY A Conjoint measurement and analysis of composition rules in psychology. Psychological Review, 78, 151-169, 1971
- 8 HAYS W L Statistics. pp484-487, London/New York, 1969

CHAPTER 6 WORKLOAD METHODOLOGY

by

Emanuel Donchin and Christopher D Wickens
Department of Psychology
University of Illinois
Champaign
Illinois 61820, USA

INTRODUCTION

The goal of our proposed technique is to employ two converging methodologies to track the workload changes during the ILS approach to landing. The two methodologies — based upon the Event-Related Brain potential (ERP) and the Sternberg Memory Search task will provide information that is both *sensitive*, detecting variations in resource demand when they occur, and *diagnostic*, localizing these changes within the multi-dimensional space underlying human processing resources (1). Each of these techniques will be briefly described.

The *Event-Related Brain Potential* is a transient series of voltage oscillations in the brain that can be recorded from the scalp in response to the occurrence of a discrete event (2). These oscillations can be characterized by a number of components which in turn may be identified by their polarity and typical latency value following the event. In a series of previous investigations (3) (4) (5) we have shown that the amplitude of the P300 component of the ERP can serve as a reliable unobtrusive index of the perceptual/cognitive load imposed by a primary task, but is insensitive to the demands associated with the selection and execution of overt responses. In this sense the measure is *diagnostic*. It has also proven to be unobtrusive in the sense that it interferes little with performance of the primary task.

The diagnosticity of the ERP dictates that it will be insensitive to variation in response load. To assess demands on the response dimension, and at the same time to confirm the variations in perceptual load that occur as the flight task proceeds, we intend to employ the Sternberg Memory Search task, also as a secondary task. This task, which has also been validated as a sensitive measure of workload in aviation environments (6) (7) requires the pilot to identify whether or not a displayed character is one of a set of characters that is held in short-term memory. Reaction time is employed to assess the speed of this decision. As the number of items in short-term memory is increased, reaction time is lengthened by an amount proportional to the speed of memory search. The time is typically indexed by the *slope* of the RT function with the increase in set size. When the central processed demands of a primary task are increased its slope increases. When the motor demands increase, the entire function shifts upward as an "intercept" increase. For example, Wickens and Derrick (8) have found intercept shifts resulting from imposing the primary task requirement to track, but slope increases as the tracking task is shifted to one involving higher order control dynamics, thereby requiring a greater amount of perceptual "lead generation".

METHODOLOGY

Both tasks will be administered as secondary tasks, by themselves and concurrently with the ILS approach scenario.

ERP During the course of the mission the pilot will hear a Bernoulli series of tones of two tone pitches, occurring at an interstimulus interval of 3 seconds. He will be asked to monitor for the occurrence of one of the tones, and make a discrete response at the fifth occurrence of each of these tones. The relevant tone will occur 33% of the time. Therefore, the discrete response will be required on the average of one every 45 seconds. On the basis of our previous research, we anticipate that each stimulus, whether relevant or not, will elicit a P300, and that P300 amplitude will covary inversely with primary task perceptual/cognitive demands. Selection of the particular interstimulus interval, relevant-stimulus probability, and stimulus modality (auditory rather than visual) is based upon our desire to choose conditions that will impose minimum interference with the primary task, yet maintain maximum workload sensitivity.

Memory Search Task Prior to the beginning of the mission the pilot will be presented a set of either 2 or 4 letters to maintain in memory. As the scenario is carried out, a series of letters will be visually presented at a prominent location on the display at an interstimulus interval varying randomly between 3 and 5 seconds. Pilots will indicate by depressing one of the two keys, mounted on the primary flight control stick whether each displayed letter is or is not a member of the designed "positive set". The latency from display presentation to response initiation will be recorded. A full replication of the workload assessment technique will require that the scenario be flown twice; once with a small memory set size ($M=2$) and once with a larger set size ($M=4$). These two replications will allow estimation of both the slope and the intercept of the function. At the present time it is assumed that the technique will employ visual presentation of letters of the alphabet. However, we are currently assessing the use of the auditory modality and a "spatial" alphabet of characters as potentially more sensitive to variance in flight related resource-demands.

DESCRIPTION OF TECHNIQUES APPLIED TO DEFINED FLIGHT TASK

Since both techniques are formally "secondary task", neither one is embedded into the primary task, and therefore neither will impose any constraints upon primary task performance as described in Appendix 1 of the original proposal. The ERP task will, of course, require auditory presentation of the probe stimuli, while the Memory Search task will require that a visual letter display device be integrated into the cockpit and a 1 bit response mechanism be available on the flight control stick.

AD-P005-633

Since the goal is to map out transient changes in workload across the mission, it will be desirable that both of our workload measures be replicated more than once at each time during the mission, so that greater reliability can be obtained. This requirement in turn dictates that the mission be flown at least twice for the ERP task and four times (twice at each memory set size) for the Sternberg Task. Our intent is to construct a running average of each measure across a sliding 10 second window, allowing this interval to dictate the "bandwidth" or temporal resolution of our measure.

LIMITATIONS AND PITFALLS

Both techniques suffer certain limitations that are inherent to some extent in the application of almost any secondary task technique. Primary among these is the issue of *reliability*: How many trials will be required to obtain an estimate of transient workload during particular points in the mission. It is difficult to establish this number a priori, but it is clear that reliability of both measures will grow with an increasing number of replications. As indicated above, we have proposed two replications as a minimum to obtain 10 second resolution.

Intrusiveness disruption of primary task performance does not appear to offer any difficulty with the ERP task. In fact this is one of its great benefits. On the other hand, the requirement for periodic overt responses in the Sternberg task may slightly disrupt performance of the continuous flight task (9).

Pilot Acceptance is a third potential limitation and it is difficult to anticipate how readily a pilot will accept (a) scalp mounted electrodes, and (b) performance of each of the associated tasks. In this regard it should be noted however that Natani and Gomer (10) and Schiflett (6) have both obtained performance measures from the ERP and Sternberg task respectively in the flight simulators and in the actual aircraft under instrument landing conditions.

REFERENCES

- 1 WICKENS C D The structure of attentional resources. In R Nickerson and R Pew (Eds), Attention and Performance VIII, Hillsdale, N J Erlbaum, 1980
- 2 DONCHIN E Surprise! ... Surprise? Psychophysiology, 18 AA AA 493-513, 1981
- 3 DONCHIN E Probing the cognitive infrastructure with event-related brain potentials. AIAA workshop on Flight Testing to Identify Pilot Workload and Pilot Dynamics, Edwards Air Force Base, California, January, 1982
- 4 ISREAL J B The event-related brain potential as an index of display monitoring workload. Human Factors, 22, 212-224 1980.
- 5 WICKENS C D Primary and secondary task analysis of step tracking: An event-related potentials approach. In R C Sugarman (Ed), Proceedings of the 25th Annual Meeting of the Human Factors Society, Rochester, New York, 1981
- 6 SCHIFLETT S G Evaluation of a pilot workload assessment device to test alternative display formats and control handling qualities. Naval Air Test Center, Final Report S7-33-R-80, July 1980
- 7 MICALIZZI J The application of factors methodology to workload assessment in a dynamic system monitoring task. University of Illinois, Engineering-Psychology Laboratory, Technical Report EPL-80-2/ONR-80-2, December 1980
- 8 WICKENS C D The processing demands of higher order manual control: Application of additive factors methodology. Engineering-Psychology Research Laboratory, University of Illinois, EPL-81-1/ONR-81-1, 1981
- 9 VIDULICH M Time-sharing manual control and memory search. Effects of input-output modality competition, priorities, and control order. Technical Report EPL-81-4/ONR-81-4 1981
- 10 NATANI K Electro cortical activity and operator workload. McDonnell Douglas Corp, Final Report MDC E2427, June 1981

CHAPTER 7

MENTAL WORKLOAD MEASUREMENT IN OPERATIONAL AIRCRAFT SYSTEMS: TWO PROMISING APPROACHES

by

Michael Biferno
McDonnell Douglas Corporation
Long Beach, California, USA

INTRODUCTION

Mental workload (MWL) is becoming a useful construct for the design of complex man-machine systems because it provides a framework to include many elements of human behaviour which are not directly observable but are vitally important in determining the safety and overall cost of the system. The non-observable aspects of a person's work include items such as remembering, interpreting, decision making and coordinating actions. In the context of highly-automated aircraft, this framework enables the manufacturer to reduce the risk of crew error and improve overall mission reliability by examining the information processing requirements of the crew to evaluate if they can perform the tasks required of them in the time available, given a clearly defined set of equipment, procedures, and operating environment.

There is a variety of techniques which have been considered as measures of MWL in aircraft systems (1)(2)(3)(4). They may be useful supplements to existing workload measures, like task-analysis which quantifies behavioural activity, when a job involves very little action but high degrees of mental activity. Although we have employed task-analytic workload measures for many years (5)(6), implementation of MWL measures in the validation of new aircraft designs has been delayed for at least three reasons. First, the requirement to systematically measure MWL has only gained acceptance in the aviation community in the last five to ten years. With higher levels of automation, the problems of overload and underload have received more serious attention by pilots, manufacturers, and regulatory agencies (7)(8).

Second, there is no agreement about the definition of MWL. This is clearly linked to the epistemological problem of dealing with an entity that cannot be directly observed. Since researchers are generally concerned with different workload applications, the evidence shows that individuals report the experience of workload with different terms (9), it is reasonable that the scientific community would formulate different definitions. The ultimate value of any scientific definition, on the other hand, rests with its "usefulness" in understanding and predicting events. To the extent that MWL maintains functional relationships with variables that are important to industry (eg, error rates, equipment costs, customer operating costs), it will be useful and cost-effective to study.

The third reason for a slow response to the problem of MWL measurement is a complex set of economic forces. On one hand industry wants the best measures available since they would reduce the risk of implementing a faulty design, on the other hand, no practical measure exists which could reduce this risk. The use of an invalid or unreliable MWL measure would induce more uncertainty of the design and certification process than would the use of current techniques. In the face of this technical uncertainty, industry can be expected to be pragmatic and employ procedures which they have found to work in the past.

Today, the manufacturers of new aircraft systems reduce their risk of designing a product which induces excessive MWL by involving company pilots (or other users) at every phase of its development. The pilots are well acquainted with operational problems and it is their responsibility to assure that a new aircraft can be operated safely by an adequately selected and trained crew. Primary methods for evaluating workload are task analytic approaches, early in design, and simulation or flight tests as the configuration firms up. Both normal and abnormal conditions are addressed.

Although the validity of this approach is strong, a number of deficiencies have been recognized by the President's task force on aircraft crew complement (8). When employing workload analyses to support future FAA certification efforts, they recommended improvements in subjective evaluation methods and a greater use of line pilots in the area of workload evaluation. Based on these recommendations we have re-examined the special requirements for MWL measurement in transport aircraft and have formulated a short and long-term plan for improving MWL measurement.

REQUIREMENTS FOR MWL MEASUREMENT

Workload measurement grew out of a requirement to insure that people could reliably perform a task when given specific equipment, procedures, operating environment, and time limitations. The techniques which are of special interest are those which can be used in high-fidelity testing so they must be noninterfering while having the necessary validity to be convincing to the design engineer, project management, user population, customer representatives and government regulatory agencies.

Applications for MWL measurement in aircraft systems can be grouped into three areas: (a) design validation or certification (eg, Can the crew member perform the assigned task?), (b) design/development aid (eg, What is the optimum crew size?), and (c) biocybernetics (eg, Can overall system reliability be improved by real-time measurement and feedback of crew MWL?). Each of these applications result in the addition of other specialized requirements. For example, if biocybernetic applications are being considered, the MWL measure must provide valid and reliable information about individuals while they are working. If an aid to design is the goal, then MWL measures should inform the design team about specific trouble spots and suggest remedial design solutions. If design validation or certification defines the requirement for MWL measurement, then the data should be suitable for answering specific questions such as: Is the new design better, worse or the same as a comparable existing system which has an acceptable safety record?

AD-P005 634

In summary, the fundamental requirements for a MWL measure are validity, reliability and usefulness. To be useful the measure must tell us something about operationally relevant types of workload in a timely and cost-effective way.

HUMAN INFORMATION PROCESSING: A FRAMEWORK

Human information processing models have been employed since the mid 1960s to organize information about human perception, decision-making and responding (10). This conceptualization of human functioning has been useful for relating the different fields of behavioural science (eg, perception, memory, and learning) and more recently, human engineering. Figure 1 illustrates the relationship between human brain structures and elements of typical human-information-processing model. Just as models based on control theory are useful in organizing the hardware elements of a system, psychological models help organize the functional elements of the human information processor (eg, receive information, decide, and act).

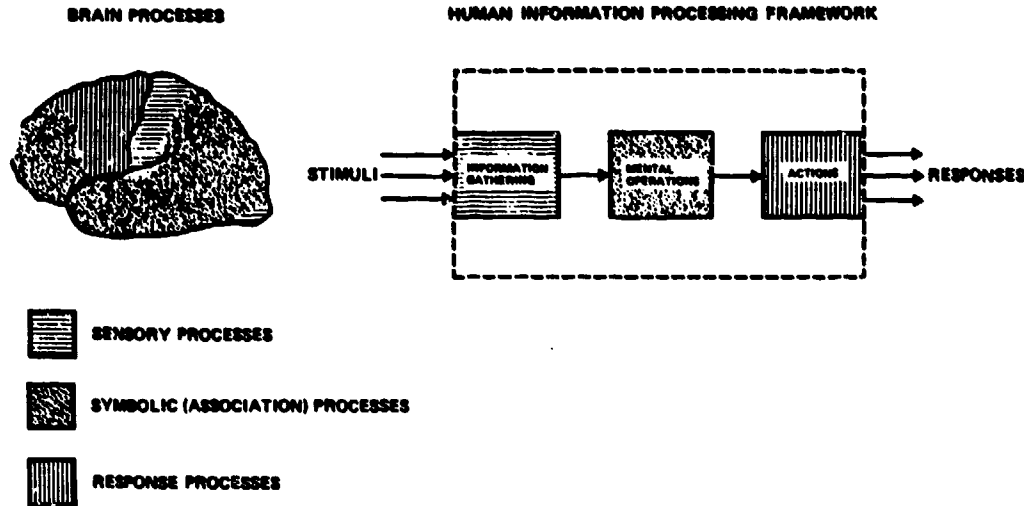


Figure 1. Relationship between Functional Brain Anatomy and Psychological Constructs in a Human Information Processing Framework.

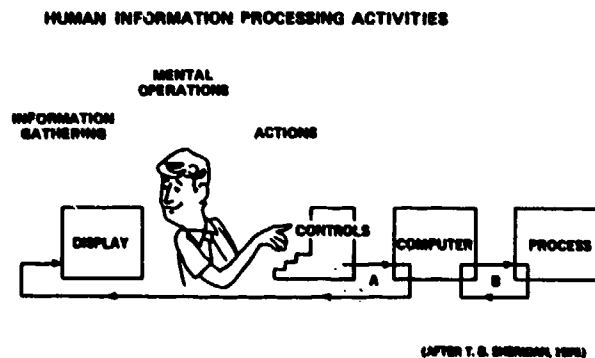


Figure 2. Human Information Processing Activities Associated with the Control of a Computer-Based Process such as Flight (after Sheridan, 1976).

This division can help clarify whether human engineering problems are associated with the transmission of sensory information (a relatively well understood area), the mental processing of information (a poorly understood area) or the performance of actions (a relatively well understood area). For example, Figure 2 illustrates a person managing a process, such as flight, with a computer-based system. The person's major duties are to gather information, interpret it and then perform the appropriate actions in a timely fashion. Starting from this framework, the sources of operator MWL can be partitioned into three areas: (a) information gathering (control of sensory processing), (b) mental operations (symbolic processing), and (c)

performing actions (control of response processing). This partitioning suggests that MWL is based in the biology of the human nervous system and the best measure may evaluate these domains separately.

MWL DEFINED AS LANGUAGE-BASED THINKING

Our definition of MWL is derived from our special requirements for design validation (certification) and development of highly-automated systems where the role of the flight crew is changing from active controller (eg, stick movement) to one of systems manager (eg, select functions and actions via system automation). Since the mental activity associated with managing a highly-automated system probably induces a high degree of verbal mediation, we have defined MWL for this application as language-based mental activity. Two methods of specifying a degree of MWL are being considered. The first views MWL as all-or-none and its degree would be expressed as a percentage of time when a crew member was experiencing a criterion level of MWL. The second views MWL as occurring along a continuum and would be expressed as a scalar which fluctuates from moment-to-moment. The assumptions underlying these two definitions are beyond the scope of this article and will not be covered, but practical considerations for a particular application may dictate the usefulness of one definition over the other.

The modern aircrew perform many language-based activities which range from preflight checklists, flight plans, and verbal communications to the management of systems. Other language-based tasks include: remembering alphanumeric data, entering data/commands into computer systems, fault isolation/problem solving, mental transformations/estimating, coordinating actions and confirming inputs. This short list is not exhaustive and does not include verbally-mediated mental sets to monitor, search, interpret, select, initiate or adjust. This conception of language-based mental activity accepts that many habits may subsume performance to non-language centers but the selection or initiation of habits probably require verbal mediation and hence, MWL.

GENERAL APPROACH

Our primary goal has been to develop measures which can be employed in crew station design and validation. Our strategy is to initiate a short-term and long-term development plan. Subjective measures are being developed for use in a two or three year time frame by adapting existing methodology where possible and implementing new techniques when necessary. Event-related brain potentials (ERPs) are being developed for applications later than three years.

SUBJECTIVE MEASURES

Background The standard practice for evaluating operator workload is to ask an expert or trained operator about the workload level associated with a particular system during a post-flight briefing. The problems associated with this type of subjective assessment are well known and yet it is the most widely employed technique because it has many attractive features, such as low-cost/ease of administration, rapid evaluation time, high face validity, data are easily interpreted, and they can be employed at every stage of development including flight testing.

A major problem with subjective measures is their susceptibility to bias. Whether the bias is due to intentional or unintentional factors, subjective measures are particularly unsatisfactory when an impartial workload analysis is required. Some of the sources of unintentional bias include: distortions of remembering, forgetting, and demand characteristics to behave or feel in predefined ways (11) (12). Sources of bias, due to a person's unique experience, can be partially controlled by evaluating multiple pilots, but the use of subjective techniques has tended to promote excessive reliance on a few experts rather than evaluate a group of representative users. Certainly, experts are needed to organize the evaluation and formulate the conclusions, but excessive reliance on a few experts can undermine the identification of problems common to pilots who are less skilled or experienced.

Ratings of workload have been available for many years but few of the techniques have established their validity and reliability with standard psychometric methods (14). Two exceptions are the bipolar-adjective rating scales developed by NASA-Ames (13) and the subjective workload assessment technique, otherwise known as SWAT developed by the United States Air Force (14). Laboratory studies have established the construct validity and test-retest reliability of these measures (15) (16).

Our experience with these rating scales indicate they are most useful when comparing the relative workload levels between two test conditions, but their application appears limited for the design and development of new equipment. On one hand the data are well suited for statistical analysis (difference tests and goodness-of-fit), but on the other hand they do not provide much insight into the underlying reasons for high workload. An analysis of the sub-scales of SWAT and the bipolar ratings, gives a clearer picture "what" the test subject experienced (eg, time-load or stress-load) not "why" the higher workload was experienced (eg, displays were difficult to read). We have found that pilots and other test subjects, are interested and cooperative in providing workload ratings but they are often frustrated in communicating the details of the equipment problems at hand. Since most rating techniques do not make provisions for a concurrent verbal report, this information must be obtained after the test when a person's memory is less accurate.

What and Why of MWL Because of the well known memory problems associated with post-flight debriefing (17), we adopted an approach which asks people to give verbal reports during the flight to supplement ratings of MWL. Put simply, our approach asks people to evaluate MWL in terms of "what and why". We ask "what" level of workload they are experiencing by means of a six-point rating scale, and if high ratings of MWL are reported, they are asked "why" (a verbal reason for the high rating). This rating technique was developed to address high MWL levels and may not be suitable for measuring the lower end of the MWL continuum (underload).

MWL is defined as the degree of attention required to perform a task. Ratings of MWL are verbalized in terms of how much attention is required to perform three basic tasks: (a) gather information, (b) perform mental operations, and (c) perform

RATE THE DEGREE OF ATTENTION REQUIRED TO PERFORM . . .

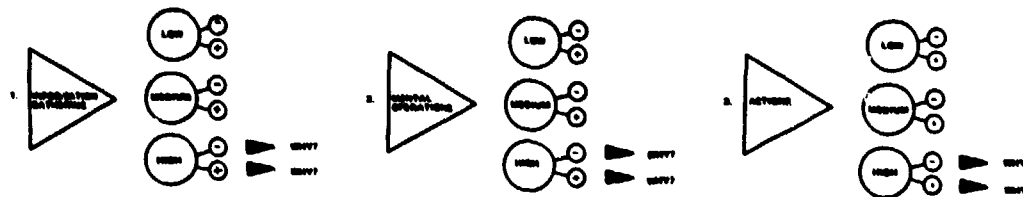


Figure 3. Three Basic Information Processing Tasks and Their Associated Rating Scales where High MWL Ratings Require a Verbal Explanation.

actions. Figure 3 illustrates the three part nature of each MWL rating with subjects rating attentional requirements as low, medium or high. Pluses and minuses are used with these labels so that subjects can refine their gross approximation with small up (+) or down (-) adjustments to yield a six point rating scale. Low, medium, and high were employed rather than a numeric scale to minimize the requirement for subjects to mentally transform their experience to a metric format.

The experience of paying attention to internal or external cues is assumed to be self-evident to subjects but they are not expected to be familiar with our method of categorizing attentional activities into three types. To ensure that subjects have a common basis for ratings of MWL, they are given examples of internal activities which require attention. Table 1 lists some activities which require attention and then organizes the activity according to three basic information-processing tasks. When subjects experience high attentional demands for any of the basic tasks, they are required to verbalize a few words or phrases indicating the cause of the high MWL. Examples for each basic task might be: "can't read the display", "radio interferes with my concentration", and "keys are hard to reach". Of course many activities can be viewed as having components of all three tasks. For example, "confirming an input" can be viewed as having sub-activities of action (moving the eye), information gathering (seeing), and mental operations (interpreting what was seen). The partitioning of the workload rating should be guided by the overriding principle of usefulness. Since the purpose of the reports is to identify MWL problem areas and associate them with controls, display or system logic, the test subjects should consider what aspect of the equipment is drawing their attention and then formulate a workload rating based on this experience.

Using this approach to review a complex system, the problem areas can be identified and prioritized by aggregating the verbal explanations of the best subjects. By assessing the magnitude of the problem across many people the problems of over-design and under-design could be avoided.

A disadvantage of this approach is the training required to administer the rating scale. The consistent use of the scale by a subject depends on a high level of motivation and sophistication. Subjects must be able to accept and use the definitions of

INFORMATION GATHERING	MENTAL OPERATIONS	PERFORM ACTIONS
<ul style="list-style-type: none"> • SEEING/HEARING/SENSING • IDENTIFYING EVENTS • IDENTIFYING OBJECTS • MONITORING • SEARCHING • RECEIVING COMMUNICATIONS 	<ul style="list-style-type: none"> • UNDERSTANDING • INTERPRETING • THINKING • REMEMBERING • CALCULATING • COMPARING • ANALYZING • PROBLEM SOLVING • PLANNING • ESTIMATING • REFLECTING INTERNALLY • SELECTING 	<ul style="list-style-type: none"> • INITIATING • TIMING • SEQUENCING • ADJUSTING • COMPLETING • CONFIRMING

Table 1. A Listing of Activities which Illustrate MWL Associated with Three Basic Human-Information Processing Tasks.

attention and apply them in a structure which distinguishes between sensing, thinking and acting. This disadvantage can be outweighed by the quality of the resulting data obtained from each test and the increase in understanding which occurs between the behavioural scientist and the technical staff who is using the data. The structure of "sense, decide and act" fit nicely into an engineering model and provide a useful heuristic for considering how the human component fits into the overall system.

It is likely that the verbal rating data obtained with this technique will be influenced by the demand characteristic to withhold high ratings until "good" justifications can be formulated. The loss of sensitivity may have positive as well as negative effects. A benefit may accrue from "filtering out" many of the reasons which are unique to a person's training, experience, and attitudes. The most practical subjective MWL measure, for engineering applications, may be one which identifies (converges on) the fundamental human information-processing problems and not the measure that identifies the greatest variety of problems (diverges). Practicality is difficult to define, but a measure which establishes agreement about the nature and priority of problems helps to move a complex design forward, whereas a measure which immobilizes the design process does not.

EVENT-RELATED POTENTIAL MEASURES

Background Electrocortical measures of workload, such as event-related potentials (ERPs), have the capability of being less susceptible to bias and less interfering than subjective measures, but a practical measure has not been developed. ERP measures can be relatively unobtrusive and noninterfering if properly implemented because they do not require conscious meditation and they can be recorded in ways which blend into many work environments (18).

A substantial amount of data supports the validity of ERPs as a measure of workload. Some experiments have found that the amplitude of the P300, a late positive component of the ERP, increases with greater workload, while others have found that it decreases. The different results depend upon the nature of the task to some degree. When the ERP is elicited by stimuli which are part of a secondary task, as primary task workload increases, P300 amplitude decreases (19) (20) (21) (22). On the other hand, when the ERP is elicited by stimuli which are part of the primary task, P300 amplitude increases as primary-task workload increases (23).

Recent work has shown that ratings of workload are related to P300 amplitude and a later component called the N400. Correlations based on individual subject data, were significant in 40 percent of the subjects tested suggesting that the relationship between ERPs and workload may be strong in some people but non-existent in others (24).

ERP Elicited by Pilot's Call-sign One way of blending ERP measurement into the work environment of a highly-automated system would be to elicit ERPs with stimuli which are part of its display devices. Since current aircraft employ digitized speech to communicate with the flight crew (eg, aural warning systems), we explored the feasibility of using a standard speech probe to elicit ERPs while people work (18). Advanced designs for future flight decks often include provisions for digitized speech to communicate with the crew. Computers can be expected to employ speech messages to annunciate changes in the automation and to alert an operator when routine tasks need to be performed. Since digitized speech can be employed across a wide variety of computer-interface applications, we devised a simple challenge-response paradigm. The computer-based system notifies the operator that their attention is required and the operator must acknowledge the request with a response, either vocal, manual or mental.

Since a speech signal reliably elicits an ERP, regardless of the subject's attentional state or sensory-receptor orientation, this approach could be employed in many work applications. Evidence from a number of tests have supported the feasibility of this approach. For example, when the subjects mentally counted the occurrence of the operator's call-sign, there was an increase in the amplitude of the late-positive potential of the ERP. This increase is compared to a condition where the subject

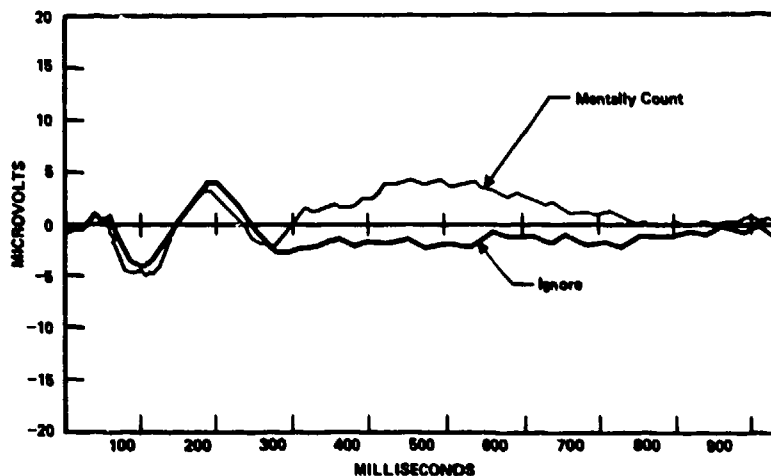


Figure 4. Increased Positive Activity Occurs after 275 ms when Subjects Mentally Count the Occurrence of Their Call-Sign (C₂ Stim).

NO GEN 241180

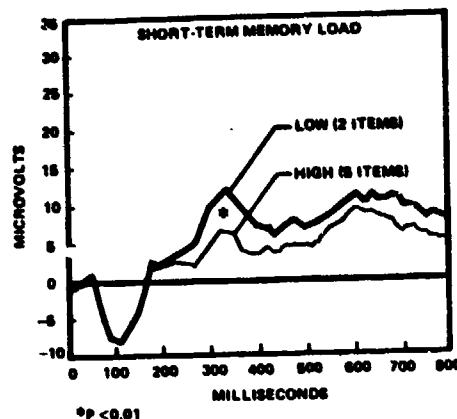


Figure 5. ERP Elicted by the Call-Sign under High and Low Levels of Short-Term Memory Load (C_z Site).

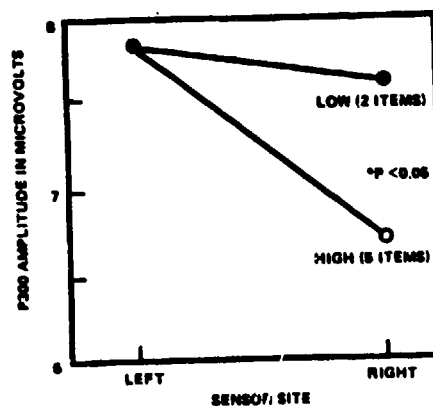


Figure 6. P300 Amplitude Elicted by the Call-Sign for High and Low Memory Load Conditions for Left and Right Hemisphere Sensor Sites.

ignored the call-sign and was engaged in another unrelated mental activity (reading a magazine). The difference in late-positive activity shown in Figure 4 can be interpreted as resulting from the mental counting activity and hence reflecting a type of language-based mental activity (25).

Although the presence of late-positive activity seems to reflect the existence of language-based mental activity, its level is reduced under some conditions of increased processing. Experimental results have shown that P300 amplitude decreases with higher memory load under a variety of conditions. When a key press is employed to acknowledge the call-sign, smaller P300s were observed at the vertex (C_z) site (see Figure 5, (26)). A slightly different outcome was obtained when a voice response (eg "Roger") acknowledged the subject's call-sign. Smaller P300s were observed over the right hemisphere but not over the left hemisphere (see figure 6, (27)).

FUTURE DIRECTIONS

Measurement of human-information-processing levels will become more accepted when the usefulness of its methods improve. The kind of methodological improvements which will have an impact on the direction of future applications are: (a) better individual-subject analysis, (b) better real-time interpretation of signals, (c) laboratory paradigms which have better generalization (external validity) to operational environments. Measurement of human-information-processing activities will become more widespread if the benefit of implementing them outweighs their costs. If a favourable cost/benefit ratio can be established, then MWL measurement can be justified on economic grounds and a firm justification can be made for its role in the long-term planning of equipment design.

SUMMARY

When evaluating aircraft systems, the most useful mental workload (MWL) measures are those which can be employed in-flight or full-mission simulations. This requires measures to be noninterfering, relatively unobtrusive, and provide estimates of operationally-relevant MWL while maintaining high levels of validity and reliability. In the context of automated systems, our strategy has been to define MWL as language-based mental activity and to develop subjective ratings (opinion scale) in the short term and event-related brain potential (ERP) measures in the long term. Subjective ratings are being employed to estimate the required degree of attention to perform: (a) information gathering, (b) mental operations, and (c) actions. This organization aids in the identification of undesirable MWL levels associated with system displays, logic, and controls. In addition to providing a quantitative workload rating, this technique elicits verbal explanations if high MWL levels are reported. The purpose of the verbal report is to identify specific items associated with high MWL ratings and to suggest alternative design solutions. The ERP is a promising objective measure which can be obtained without interfering with normal work activities regardless of the crew's sensory-receptor orientation or conscious state. Experimental results support the feasibility of this approach by using stimuli already present in the modern cockpit (digitized speech) to elicit ERPs that change with increased memory load and workload ratings.

REFERENCES

- 1 HARTMAN B O
McKENZIE R E (Eds) Survey of methods to assess workload. AGARD-AG-246, available through National Technical Information Service, Springfield, Virginia, 1979
- 2 MORAY N (Ed) Mental Workload: Its theory and measurement. New York: Plenum. 1979
- 3 WIERWILLE W W Physiological measures of aircrew mental workload. Human Factors, 21, 575-993 1979
- 4 WILLIGES R C
WIERWILLE W W Behavioural measures of aircrew mental workload. Human Factors, 21 549-574 1979
- 5 STONE G
REGIS E R
GULICK R J DC-9 Super 80/DC-9-50 comparative flight crew workload study. MDC-J8749, Douglas Aircraft Company, Long Beach, California 1980
- 6 BROWN E L
STONE G
PEARCE W E Improving cockpits through crew workload measurement. MDC-6344, Douglas Aircraft Company, Long Beach, California 1975
- 7 HAY G C
HOUSE C D
SULTZER R L Summary report of 1977-1978 task force on crew workload. FAA-EM-78-15 Federal Aviation Administration, Washington DC 1978
- 8 ANON Report of the president's task force on aircraft crew complement. Department of Transportation. 2 July 1981, Washington DC 1981
- 9 HART S G
CHILDRESS M E
HAUSER J R Individual definitions of the term "workload". Paper presented at the symposium: Psychology in the Department of Defense, US Air Force Academy 1982
- 10 REITMAN W Cognition and thought: an information processing approach. New York: Plenum 1965
- 11 STAVE A M The effects of cockpit environment on long-term pilot performance. Human Factors, 19, 503-514 1977
- 12 WALSTER B
ARONSON E the effects of expectancy of task duration on the experience of fatigue. Journal of Experimental Social Psychology, 3, 41-46 1967
- 13 HART G S
BATTISTE V
LESTER P T POPCORN: A supervisory control simulator for workload and performance research. Proceedings of the 20th Annual Conference on Manual Control 1984
- 14 REID G B
EGGEMEIER F T
SHINGLEDECKER C A Application of conjoint measurement to workload scale development. Proceedings of the Human Factors Society 25th Annual Meeting, 522-526 1981
- 15 CHILDRESS M E
HART S G
BORTOLUSSI M R The reliability and validity of flight task workload ratings. Proceedings of the Human Factors Society 26th Annual Meeting, 319-324 1982
- 16 EGGEMEIER F T
et al Subject workload assessment in a memory update task. Proceedings of the Human Factors Society 26th Annual Meeting, 643-647 1982
- 17 REHMANN J T
STEIN E S
ROSENBERG B L Subjective pilot workload assessment. Human Factors, 25, 297-307 1983
- 18 BIFERNO M A
BIGHAM T R Speech-related potentials elicited by synthetic speech stimuli. Psychophysiology, 19, 306-307 1982

- 19 ISREAL J B
et al P300 and tracking difficulty: Evidence for multiple resources and dual-task performance. Psychophysiology, 17, 259-273 1980
- 20 ISREAL J B
et al The event-related brain potential as an index of display-monitoring workload. Human Factors, 22, 212-224 1980
- 21 NATANIK
GOMER F E Electroocortical activity and operator workload: a comparison of changes in the electroencephalogram and in event-related potentials. McDonnell Douglas Corporation, St Louis, Missouri, McDonnell Douglas report MDC-E2427 1981
- 22 BIFERNO M A Short-term memory influences on event-related potential components and lateralization. Unpublished manuscript, McDonnell Douglas Corporation, Douglas Aircraft Company, Long Beach, California 1985
- 23 HORST R L
MUNSON R C
RUCHKIN D S Event-related potential indices of workload in a single task paradigm. Proceedings of the Human Factors Society 28th Annual Meeting, 727-731 1984
- 24 BIFERNO M A Mental Workload Measurement: Event-related potentials and ratings of workload and fatigue. NASA report number CR-177354, Ames Research Center, Moffett Field, California 1985
- 25 BIFERNO M A N200 latency in event-related potentials elicited by tone and synthetic speech stimuli. Psychophysiology, 19, 551 1982
- 26 HILTON P G The influence of mental workload on decision-making time and subjective workload assessments. Master's Thesis, California State University, Long Beach, California 1983
- 27 NORDENSTAM B The effect of increased memory load on the event-related potential and hemispheric lateralization. Master's Thesis, California State University, Long Beach, California 1985

CHAPTER 8

CORTICAL EVOKED RESPONSE AND EYEBLINK MEASURES IN THE WORKLOAD EVALUATION OF
ALTERNATIVE LANDING SYSTEM DISPLAYS

by

R D O'Donnell
Ergometrics Technology, Inc
4401 Dayton-Xenia Road
Dayton, Ohio 45432, USA

and

Glenn Wilson
Harry G Armstrong Aerospace Research Laboratory
Wright-Patterson Air Force Base
Dayton, Ohio 45433, USA

INTRODUCTION

From an intuitive viewpoint, physiological measures would appear to provide an optimal set of techniques for assessing workload. They make minimal demands on the operator's time and attention, they lend themselves to ready quantification, and they tap functions which are easy to relate theoretically to the workload construct. For example, in one view (1) a major determinant of workload is the amount of effort required of the operator. It would be expected rationally that the amount of effort expended should manifest itself in the degree of physiological arousal or activation in the individual. Therefore, indices of such arousal should bear a direct and consistent relationship to the amount of workload.

Unfortunately, the history of attempts to derive such direct relationships is less than impressive. While there have been notable successes, such as the work of Beatty (2) on pupillary measures and the heart rate results reported by Roscoe in this volume, there were many instances in which physiological measures failed to show correlations either with imposed workload levels, or even with each other under identical conditions. Such results led some investigators to abandon physiological measurement entirely on the basis that it was inherently unreliable. Others, however, realized that such lack of correlation might just as well indicate that the measures were tapping different aspects of a complex construct and might, in fact, be revealing an unexpected and very desirable specificity or "diagnosticity". Combined with the "global" indices of activation represented by such things as pupil diameter and heart rate, highly diagnostic physiological measures might pinpoint the type of processing resource, stage, or strategy which is being loaded by a particular task. This realization came at a time when workload theorists were emphasizing the multi-dimensional nature of the workload construct. Thus, there is a happy correspondence between the need in workload assessment for measures which tap specific resources or stages, and the growing realization that some physiological measures may be quite specific in their sensitivity to precisely such types of psychological function.

Based on the results of a number of studies the U.S. Air Force decided, in 1979, to construct a battery of physiological tests, each of which had shown some promise in laboratory studies of being sensitive to various aspects of workload. This Neuropsychological Workload Test Battery (NWTB) is currently undergoing validation testing in several simulator environments. Two of the most promising measures from this battery are the transient cortical evoked response and several analyses of eyeblink behaviour. It is becoming clear that these techniques can contribute complementary types of information on the amount of workload being experienced by the operator, and could form the basis of a measurement system which would tap both global and specific aspects.

RATIONALE FOR THE MEASURES

The transient evoked response is obtained from the electroencephalogram (EEG) when a discrete stimulus is presented to the subject in some sensory modality. In order to isolate this response from the ongoing EEG activity, multiple stimulus presentations are usually necessary, and the brain activity following the stimuli are time-locked averaged to enhance the signal-to-noise ratio. The typical response can be variable between individuals with respect to precise amplitude and latency of peaks, but usually shows the same general morphology. This consists of two positive peaks prior to about 250 milliseconds after the stimulus, and one major peak between about 250 and 500 milliseconds. This latter peak (called the P3 or P300 peak) is found only when the subject is actively processing information, and when stimuli have some relevance to the task being performed by the subject. In addition, within an individual, the amplitude and latency of the P300 appear sensitive to different aspects of the task. Amplitude appears to be directly proportional to the degree of subjective surprise at the appearance of the stimulus, whereas latency appears to vary with stimulus evaluation time (3). Further, under certain conditions, it has been determined that when the stimuli are presented during performance of a "primary" task (which may be visual, auditory, visual-motor, etc.) the amplitude and latency of the P300 show a remarkable sensitivity to the workload of the primary task (4).

Further studies indicated that this sensitivity was specific to the perceptual/central processing demands of the task, and was insensitive to the motor demands (5) (6). Thus, as used by these authors and adopted in the NWTB, the transient evoked response can be viewed as a highly diagnostic measure of the central processing (mental) workload of the operator. It would appear to be particularly appropriate in those situations where the operational environment makes it difficult or impossible to obtain objective, short-term measures of the amount of mental activity required by a task.

In a similar way, it is known that the blink pattern of individuals changes as a function of several aspects of task demands, as well as of subject state. However, studies which utilized blink frequency as a dependent variable have been severely criticized because of problems in design, analysis or experimental control (7). Such measures appear to show great variability, and require a degree of experimental control which would tend to preclude their use in operational settings.

Other approaches to the analysis of eye blinks have been considerably more successful in assessing longer-term effects of workload (8). Studies have indicated that humans have characteristic patterns of blink behaviour which are quite specific to the task, and which are altered only under conditions of stress, fatigue, or task load.

DESCRIPTION OF THE TECHNIQUES

The Evoked Response Oddball Paradigm. A particularly powerful technique for obtaining the transient evoked response in a workload situation has been reported by the Cognitive Psychophysiology Laboratory of the University of Illinois (3). In this procedure, called the "oddball" paradigm, the subject is required to attend to "secondary" stimuli during the performance of a "primary" task. Typically, the secondary stimuli are of two easily discriminable classes (e.g., two tones of different frequencies), one of which occurs much more often than the other. For instance, while the subject is tracking a visual target (the primary task) high tones may be presented through earphones 80% of the time, and low tones may be presented 20% of the time. The subject is instructed to monitor (e.g., to count) either class of tones (the secondary task) while performing the tracking task. The sequence of tones may be random or controlled (e.g., presented in a Bernoulli series) depending on the specific goal of the experiment. In any case, the evoked response generated by one or more of the "rare" stimuli (32 to 64 stimuli are frequently used) is obtained, and the P300 amplitude and latency are determined. These values are then used to index the central processing workload of a task.

For this measure, recording electrodes are attached to the scalp. Although the precise placement is not critical for the recording of the P300 wave in this paradigm, records are usually obtained from standard left and right parietal leads, referenced to linked mastoids. As with most physiological measurement, it is important to reassure the subject that the procedure is harmless and non-invasive. Given this, subjects typically have responded extremely well to this type of measurement, and rapidly forget about the electrodes and the data acquisition. Further face validity can be introduced into the test by utilizing tones that naturally occur in the environment (radio signals, threat warnings, normal environmental sounds) as the stimuli. With proper attention to the requirements of the paradigm, utilization of such naturally occurring stimuli can significantly enhance the co-operation from the subject and the overall validity of the data.

Eye Blink Recording Eye blinks can be recorded with an eye point of regard monitor (see chapter 9) or by means of some form of electro-oculogram (EOG) using miniature electrodes placed above and below an eye (8).

EXAMPLE OF USING THE TECHNIQUE

Recent advances in aircraft landing systems (e.g., microwave landing systems) permit several new control strategies to be introduced. For instance, complex landing paths can now be directed from ground control stations. However, such innovations require that flight directors be redesigned, and that the data they provide to the pilot be changed from that currently used. Several new designs are available, and among the many questions that must be answered before one can be chosen, determination of the workload of each system is one of the most critical. Assuming that overall aircraft performance is relatively equal for all systems (a necessary prior determination before workload even becomes an issue) it is required that a quantified ordering of the workload involved in each of three systems be produced.

The mission consists of several complex approaches in a large aircraft simulator. Each approach lasts for 10 to 20 minutes, and the various approaches involve different specific tasks (e.g., number of turns) as well as different environmental factors (e.g., turbulence). Subjects will consist of ten volunteer line pilots.

In this situation, it is necessary to realize that a baseline measure is absolutely essential. As in behavioral secondary task measures (see Shingledecker, this volume) physiological measures such as the evoked response and eyeblinks must be interpreted in terms of a stable individual baseline. Therefore, the first requirement is to establish a comparable set of approaches with existing, known flight directors having a previous history of performance acceptability. If all approaches cannot be baseline, at least a moderately complex one should be obtained.

A second procedural technique is also desirable in order to help establish the "operational" meaning of any physiological differences found between systems. One of the most frequent errors in physiological studies in operational environments is to stop when statistically significant differences are discovered between experimental conditions. While these may be theoretically important, and may often be interpreted in terms of the construct validity of the test, they almost always fail to convince the operational pilot that the differences have any practical value. To help alleviate this, it is desirable to introduce several levels of stress *within* each of the experimental conditions. In this way, changes for each landing system as a result of constant stress increases can be compared. If these functions are different for one system with respect to another, or with respect to baseline conditions, even the non-physiologist can see that the systems differ in their workload.

The final design would then be a flight director by approach type by stress level factorial. Given four flight directors (including the control condition), three approach types, and two stress levels, each subject would fly 24 simulator missions. A full battery of primary task performance measures such as glide slope error, etc., would be taken. Several workload measures would also be obtained, designed to sample the global workload, and to probe specific components of the workload construct in order to diagnose the source of any workload "chokepoints" identified by the global measures.

The cortical evoked response would be useful in this case to obtain relatively short-term estimates of the central processing load introduced by the different flight directors. Auditory tones, made to simulate normal tones occurring in the cockpit, would be presented randomly but with different probability of occurrence. The tones would require attention, but no

response, from the pilot. For previously defined segments of the approach (lasting about 3 to 4 minutes each) the transient response to these tones would be obtained. Segments would be chosen to be representative of a range of workloads in the overall approaches.

From this design, 3 to 5 evoked responses per mission will be obtained. Differences within each subject (delta scores) between baseline mission segments and comparable segments within each of the other missions will be used as the primary data. These delta scores can then be compared between flight directors for each subject, and composite statistics generated by appropriate techniques. Similarly, the delta scores can be used to generate a "workload increase with a given stress" for each flight director and for the control condition. This derivative function will further serve as a sensitive measure of subtle workload differences between experimental conditions.

Eyeblink measures will be used in the present case to obtain both global and specific indicators. The histograms of interblink intervals will provide a time history of differences between flight directors in such things as stress, information demand, and workload. Analyses will be similar to those described above, with delta scores based on the subject's baseline condition and response to stress in that condition used as raw data. However, the time period covered by one histogram will be somewhat shorter (about one minute), allowing somewhat finer resolution of the workload history. Since this is a time-based measure rather than an event-based measure, it is more important that the events occurring during each time period be examined and related to the results. Further, since the blink measure is not able to differentiate easily between central and motor load, inspection and correlation of the events occurring in the scenario is even more important to interpretation.

Similar analyses can be carried out for closure duration. In addition, however, one is particularly interested in the occurrence of "atypical" closure durations, either unusually long or unusually short. In the one case, long duration closures indicate the occurrence of "dropouts" in performance — a very significant event if it happens in an aircraft. In the other case, very short closures occur under high information load and stress, and would indicate an undesirable situation if continued for a long time.

PITFALLS AND LIMITATIONS

The basic pitfall in the use of any physiological measurement is the difficulty of keeping both the experimenter and the subject from becoming either too enthusiastic or too disappointed with the techniques. Neither extreme is justified. Physiological metrics provide a valuable adjunct, and nothing more or less, to subjective, behavioral, and modelling techniques for assessing workload. Whether they provide redundant information, ancillary information, or information which can be obtained in no other way depends entirely on the question being asked and the environment in which it must be answered.

With respect to the specific techniques described here, several cautions must be clearly pointed out. The evoked response is usually obtained as an *averaged* phenomenon. Therefore, very short-term changes not only cannot be detected by the procedure, but may actually confound the average. Therefore, it is critical that the stimuli used to generate the evoked response in the oddball paradigm be, as nearly as possible, equal in relevance to the subject over the entire data collection period. Lacking this, consideration should be given to single trial evoked response techniques.

From a very practical viewpoint, it must be realized that the evoked response is a small electrical signal. While electrodes and amplifiers have progressed to the point where normal movement or electrically noisy environments are not insurmountable problems, they must still be considered and controlled. Artifacts occur in the data under the best of circumstances, and experimenters should be trained to detect and eliminate them. In the same way, identification of peaks in the evoked response is not always straightforward. Although it is usually unambiguous in most subjects, there are enough anomalous cases that a trained observer is still needed unless very sophisticated computer software is available.

For eyeblink recording, the limitations are similar to those for evoked responses. While the signal is electrically somewhat larger, the eye tends to do more things to interfere with the desired response — the subject blinks, squints, raises the eyebrows, and twitches. All of these may be difficult to discriminate from the eyeblink in the normal EOG, especially for the untrained observer. For this reason, it is usually necessary to "screen" the records and eliminate artifacts before processing. Thus, this measure also requires a trained analyst, even if the analysis is fully automated after the initial screening.

Head movements can be particularly disturbing in this technique, especially if they involve large movements pivoted on the neck (such as might be seen in fighter pilots). In severe cases, it might be necessary to use accelerometers attached to the head to reveal such movements so that they could be screened out from the eye movement record. Large head movements where the body accounts for the majority of the movement (e.g., bending or twisting movements) do not appear to cause severe recording problems.

SUMMARY

Physiological measures will be of use in the assessment of workload to the extent that researchers attend to the validated global or specific diagnosticity of measure. They should be viewed neither as a panacea, nor as a frivolous add-on to an experiment. Specifically, the cortical evoked response and the analysis of eyeblink behaviour can provide both global and specific indicators of workload, and when used with appropriate caution, can yield valid measures in situations where other objective measures are difficult or impossible.

REFERENCES

- 1 JOHANSSON G
et al Final report of the experimental psychology group. In: Moray N (Ed) Mental Workload: Its theory and measurement. Plenum Press, New York, 1979.
- 2 BEATTY J Task-evoked pupillary responses, processing load, and the structure of processing resources. Psychological Bulletin 91 (2) 276-292 1982.
- 3 DONCHIN E Event-related brain potentials: A tool in the study of human information processing. In: Begleiter H (Ed) Evoked potentials in psychiatry. Plenum Press, New York, 1981.
- 4 WICKENS C D
ISREAL J
DONCHIN E The event-related cortical potential as an index of task workload. Proceedings of the 21st Annual Meeting of the Human Factors Society, San Francisco, 1977.
- 5 ISREAL J
WICKENS C D
DONCHIN E The event-related brain potential as a selective index of display load. Proceedings of the 23rd Annual Meeting of the Human Factors Society, Boston, 1979.
- 6 ISREAL J
et al The event-related brain potential as an index of display monitoring workload. Human Factors, 22 211-244, 1980.
- 7 HALL R J
CUSACK B L The measurement of eye behaviour: Critical and selected reviews of voluntary eye movements and blinking. US Army Technical Memorandum 18-72 Human Engineering Laboratory. Aberdeen Proving Ground, Maryland, 1972.
- 8 OSTER P J
STERN J A Measurement of eye movement. Electro-oculography. In: Martin I and Venables P H (Eds) Techniques in psycho-physiology. J Wiley & Sons, New York, 1980.

CHAPTER 9

IN-FLIGHT ASSESSMENT OF WORKLOAD USING INSTRUMENT SCAN

by

J R Tole
Digital Analysis Corp
Box 2850, Reston, Virginia 22090, USA

and

R L Harris Sr
NASA Langley Research Center.
Hampton, Virginia, USA

INTRODUCTION

During instrument flight, the pilot obtains information concerning aircraft state by cross-checking or scanning the flight instruments. The exact method of scanning the instrument panel varies from pilot to pilot but there are some basic features common to a "good" scan pattern. Indeed, it was the early study by Fitts and his associates identifying the most common instrument transitions which led to the familiar "T" arrangement of the major flight instruments (1).

The method discussed here may be considered a candidate for workload studies with piloting tasks which will invoke a regular visual scan (spatial/temporal pattern of eye movements) during instrument flight. When instrument scan is in use, it may be postulated that external factors such as noise, interruptions, fatigue, etc which interfere with the piloting task may produce measurable changes in the scanning behavior. Such measures would be particularly attractive for quantifying workload since they would be both non-invasive and objective.

It is important to point out that instrument scan by itself is not a complete indicator of workload nor is task attention necessarily associated with where the pilot happens to be looking at a particular instant. However, whenever instrument scan is required in a piloting task, analysis of scanning behavior may yield important direct or indirect information concerning workload.

Scenarios in which instrument scan may be considered a potential candidate for workload assessment include:

- 1 Any situation in which instrument flight is required as part of the overall task,
- 2 Alterations in the design and/or layout of cockpit instruments.
- 3 Alterations in controls which require visual monitoring of.
- 4 Situations in which fatigue is suspected to be high.

METHODOLOGY

Measuring Visual Scan

Measurement of pilot lookpoint (eye point-of-regard) is required in order to analyze the instrument scan. While several techniques have been applied over the years, the most practical method for in-flight measurements is the remote oculometer. This device makes no contact with the pilot and does not restrict his movements while tracking his point of regard to within approximate 0.5 degree accuracy. The oculometer measures infrared light (from a low intensity source in a corner of the instrument panel) reflected from the retina and cornea of the eye via an infrared sensitive TV camera and a system of lenses and mirrors. Computer analysis of these reflections is performed to determine where the pilot is looking. Basic output from the oculometer consists of the x, y coordinates of the visual scene as a function of time. Temporal resolution is 1/30 second. For convenience in later analyses, the raw data is usually converted to yield instrument dwell rather than the x-y coordinates.

NASA Langley Research Center has devoted considerable effort to the problem of installing such a device in a cockpit. The current version of their electro-optical (EO) head requires a little more than the space of an instrument on the instrument panel. A more complete description of the oculometer is available elsewhere (2).

Analyses of Scanning Behavior

Analyses of the information provided from the oculometer may be separated into temporal, spatial, and spatio-temporal categories. In all cases, the fundamental premise is that the 'regular' scan path will in some way be altered by some factor(s) (eg panel layout) which may affect workload during instrument flight. The analyses described here do not by themselves measure workload, however they allow comparisons of the scan path behavior of the pilot under various situations and thus may provide inferences concerning changes in workload.

Temporal Analyses

Time History of Lookpoint

The fundamental output from the oculometer is a time history of lookpoint (ie a plot of the instrument being viewed as a function of time). Besides providing the basic data from which other analyses may be performed this plot is useful as an overview of the scanning behavior; eg it is particularly easy to determine periods of 'staring' or high rates of blinking.

AD-P005 636

Dwell Percentages

The dwell percentage is the percentage of time spent looking at a particular instrument. The transition percentage is the percentage of transitions which occurred between two instruments regardless of the direction of the transition. These data are printed on a schematic view of the instrument panel with the dwell percentages inside the individual instrument boundary and lines between the instruments representing those transitions which occurred (the width of the line can be drawn proportional to the magnitude of the transition percentage). This diagram gives a graphic picture of the scan paths.

Dwell Histograms

Dwell time histograms may be plotted for each of the important instruments. Such a histogram is a plot of the number of dwells (looks) on an instrument which lasted for the length of time indicated by the abscissa. Intuition suggests that instruments with either high information content or poorer information transferability will elicit longer dwells than those with low amounts of information or good information transferability. When additional information is added to a display or the display format is changed, dwell histograms may be successfully used to examine the effect of this change on the pilot (2). The goal is to arrive at a display design which will provide the most information with the shortest dwell time.

Dwell time histograms tend to be stereotyped in shape for different instruments. Dwells can be classified by both the instrument being looked at as well as the function of the dwell, i.e. whether the pilot was monitoring information or changing the indication by some control input while looking at the instrument. The histograms for these two dwell functions have two peaks, one at short dwell times for 'check' on aircraft state and a second peak at longer dwell times associated with the 'reading' of aircraft state. The control dwells show a peak at a very long dwell time (2).

Oculomotor Dynamics

Oculomotor dynamics are a useful type of ancillary data which may be considered during scan path analyses. While not a direct indication of scanning behavior, the details of how the eye moves between instruments may be an important indicator of fatigue. In particular, peak velocity and acceleration of saccadic eye movements can be expected to decrease dramatically as the oculomotor system fatigues. Measurement of these parameters can provide an indication of the tendency to fatigue under certain types of instrument scan.

Spatial Analyses

Instrument Transitions

The earliest analyses of the instrument scan calculated the probabilities of a pilot making a change in lookpoint between pairs of flight instruments. The instrument transition matrix results from determining the probabilities of all such changes which are possible. While it is theoretically possible to statistically compare two such matrices, obtained under different workload conditions, the amount of data required to make such a comparison valid is often more than can be obtained in a practical situation. This fact led to the development of a single parameter measure of scan behavior, called entropy, which in effect summarizes the probabilities contained in the transition matrix (3).

Entropy

The time history of fixations has a form which is similar to that of a communication system which can assume N discrete states with a varying duration in each state. The orderliness of such a system is related to the probabilities with which it occupies its different states. A system which always occupied the same state or always made the same transitions between states would thus be quite orderly. In the case of instrument scan, these situations would be paralleled by staring and by a stereotyped scanpath respectively.

This concept of system order may be stated compactly (4) as:

$$H_o = - \sum_{i=1}^D [p_i \log_2 p_i]$$

where H_o = observed average entropy

p_i = probability of sequence i occurring

D = Number of different sequences in the scan

In the case of the instrument scan, entropy has the units of bits/sequence and provides a measure of the randomness (or orderliness) of the scanpath. The higher the entropy, the more disorder is present in the scan. The maximum possible entropy is constrained by the experimental conditions. The maximum possible value, H_{max} , may be calculated as follows. For a given number of instruments, M , and sequence length N , the maximum number of different fixation sequences is given by:

$$Q = M * (M-1)^{N-1} = \text{maximum of sequences of length } N$$

The number of bits required to uniquely encode all Q possible sequences is $\log_2 Q$. The magnitude of this latter number also represents H_{max} of the visual scan for the number of instruments and sequence length being considered. For example, with 7 instruments the value of Q for sequences of 2 instruments is 56 which yields a corresponding $H_{max} = 5.8$.

In order to include the effect of instrument dwell times, a term for entropy rate may be defined as:

$$H_{rate} = \sum_{i=1}^D [H_i / DT_i]$$

where H_i = entropy for i th sequence
 DT_i = Average dwell time for i th sequence
 D = Number of different fixation sequences

While it is possible for pilots to make rather rapid glances (with dwell times of 100 msec or less) at their instruments (5) a fixation rate this high (10 fixations/sec) rapidly leads to oculomotor fatigue. A more realistic average value is probably about 2 fixations/sec or less for a long period of instrument scan (say > 10 sec). Using this value (0.5 sec/look) as the average dwell interval, the maximum entropy rate for 7 instruments and sequences of length 2 is calculated from the following equation to be:

$$(H \text{ rate})_{\max} = 5.8/0.5 * 2 \text{ fixations/sec} = 6 \text{ bits/sec}$$

This number represents an upper bound. Since we suspect that the pilot must have some regularity in his or her scan, the numbers we would expect to obtain under actual flight conditions will probably be lower. The observed average rate for the basic experiments was on the order of 1 bit/sec. A tendency to stare under increased load should be reflected by decreased entropy and increased fixation times making H rate tend toward lower values under such conditions.

Spatio-temporal Analyses

Correlation

In situations in which a workload inducing stimulus is applied either periodically (eg verbal loading, secondary task, etc) or in a recurring but random fashion, the use of correlation methods may be in order.

Autocorrelation may be performed on scanning data as follows. A sequence of instrument numbers versus time is developed from the data. Due to the arbitrary nature of the assignment of instrument numbers, the autocorrelation of the signal containing all instrument numbers does not necessarily produce meaningful results. For this reason analysis of each instrument is examined successively by replacing the time sequence of all instruments with a sequence $[x(i)]$ where the value is 1 for the instrument being studied and 0 for all other instruments. In order to eliminate the dc component for later spectral analysis, a zero-mean sequence $[f(i)]$ is computed from $[x(i)]$ as follows:

$$f(i) = x(i) - \bar{x}_i$$

where

$$x_j(i) = 1$$

if specified instrument j is being fixated and 0 otherwise

$$\bar{x}_j = \text{mean of } [x_j(i)]$$

The sample autocorrelation of $[f(i)]$, or sample autocovariance of $[x_j(i)]$, is calculated by the formula:

$$R_j(k) = 1/n \sum_{i=1}^n [f_j(i) * f_j(i+k)]$$

where $R_j(k)$ = autocorrelation sequence for instrument j

n = number of samples = total run duration/oculometer sampling period (1/30th sec)

This autocorrelation is computed for each instrument for each loading case. In order to detect possible periodicity in the scan, the Fourier transform of the autocorrelation is taken to produce the power density spectrum. From this a value for the dominant frequency may be obtained. For skilled pilots, this frequency tends to be close to that of the workload stimulus which has been applied. This suggests that the pilot has a tendency to multiplex the flying task and the periodic task for greater efficiency. Overload occurs when numbers are presented too rapidly for the pilot to efficiently multiplex both tasks.

Novice pilots, however, do not seem to have any consistent pattern in their autocorrelation sequences. Most of these pilots show little or no periodicity in their scans for any of the loading conditions. One explanation may be that skilled pilots have a better developed ability to time multiplex several simultaneous tasks.

For stimuli which occur repetitively, but not at periodic intervals, it is plausible to consider the use of cross correlation between the time at which the stimuli are applied and the scanpath although this has not been attempted to date.

Visual Scanning Measures Applied to the Standard Flight Task, (Manually Flown ILS Approach and Landing of Two-Pilot Passenger Jet Transport).

We now briefly discuss the application of our techniques to the valuation of workload during an ILS approach. Two or three factors must be manipulated to use the techniques described above: (a) a piloting task requiring a stereotyped scan path, (b) a verbally presented mental loading task, or (c) a visually presented mental loading task. It is assumed that the cockpit to be used for the experiments may be outfitted with the NASA Langley oculometer system or an equivalent and that ample time will be allowed (approximately 5-10 minutes) for calibration of the oculometer before an experimental session begins.

The proposed ILS approach scenario requires the use of a stereotyped scanpath, though it should be emphasized that the task and hence the scan pattern is not constant throughout the scenario. Thus, the second to second level of loading due to the flight task and the corresponding instrument scan will vary, albeit in a somewhat predictable fashion. The additional verbal or visual loading task serves to "bias" the total amount of mental load on the pilot with the goal of locating peaks in the load due to the piloting task alone. The notion here is that the workload due to the additional task is roughly additive with the instantaneous load due to the piloting task. The hope would be to bias the total load to a high enough level to demonstrate a performance decrement (which may be a non-linear function of loading) while at the same time hopefully observing a monotonic change in the measures of scanning behavior as a function of the increased load.

Several levels of difficulty of the additional task are required. These may be achieved in two ways. A constant level of difficulty may be imposed over the entire approach; this method is to be recommended at present as we are not as yet sure how to analyse short segments of the scan pattern. Each level of difficulty of the imposed extra task would thus require a separate run. Since both the verbal and visual tasks are periodic, their respective difficulties may be altered during a run by changing the period between presentations of the task. This method would seem more attractive if the piloting task were indeed fixed over the entire run.

A verbal task may be used as one means of biasing the loading level. This has been shown to work well in our experiments and is easy to implement and score (6). Such a task should be designed to approximate one which would ordinarily be performed in the course of flight; eg a constant rate of radio communication or periodic manual computation of navigational coordinates.

An alternate, visual version of this task is also possible and perhaps more appropriate for actual flight conditions. A small display could be mounted in a convenient point in the pilot's visual field. The display could present either a "+" or a "-" sign. At periodic intervals an auditory "beep" would signal that the pilot should observe this display and indicate (operationally) via a rocker switch whether the display is currently indicating + or -. The interval between "beep" determines the difficulty of this task and one possible measure of workload is the % of time the pilot is actually able to observe the display.

Entropy rate calculations could be made on the scanning data regardless of whether the visual or verbal loading task is used. Since both tasks are periodic, the autocorrelation technique may also be applied. Although we have not done it as yet, we expect that cross correlating the time of presentation of the imposed task with the scanning data is likely to yield good results especially in the type of flight scenario proposed in this study. We expect that a characteristic "signature" will appear in the cross correlation between the loading task and the instrument scan and that this signature will be altered via changes in task difficulty.

Limitations and Pitfalls of the Technique

There are a number of potential problems in applying our techniques. These are enumerated below:

- 1 The piloting task being performed must require instrument scan.
- 2 The relationship between where the pilot is looking and the 'focus' of his attention may be misleading (clearly this is the case if the pilot is staring).
- 3 The scan must be repetitive, at present, although it may be possible (eg using cross correlation) to analyze short segments of a scan pattern.
- 4 An onboard oculometer is required and must be mounted in the instrument panel (NASA — Langley Research Center has worked out many of the technical problems however). Jet Transport simulators at NASA Langley and elsewhere have also been fitted with the oculometer.
- 5 It may be necessary to calibrate without the pilot's cooperation due to time limitations in the proposed experiments. However sufficient setup time prior to the experiment will minimize the calibration needed.
- 6 The behavior of the various measures of scan has not been examined under a wide variety of situations as yet, hence we are unable to comment on flight scenarios in which the task is most applicable other than the obvious requirement of some type of scanning behavior.

REFERENCES

- 1 JONES R E
MILTON J L
FITTS P M Eye fixations of aircraft pilots: A review of prior eye movement studies and a description of a technique for frequency, duration, and sequence of eye fixations during instrument flight. USAF Tech Report 5837 (AT I 65996), 1946
- 2 HARRIS R L Sr
GLOVER B J
SPADY A A Summary of analysis techniques of pilot scanning behavior and their application, NASA Langley Research Center, Technical Report, 1985
- 3 TOLE J R
et al Entropy, instrument scan and pilot workload. IEEE Conference Proceedings Systems, Man, and Cybernetics, 1982
- 4 MIDDLETON D B
HURT G J Description and flight tests of an oculometer. NASA Technical Note NASA TN D-8419, June 1977
- 5 HARRIS R L Sr
CHRISTHILF D What do pilots see in displays? presented at the Human Factors Society Meeting, Los Angeles, October 1980
- 6 STEPHENS A T Instrument scan, performance, and mental workload in aircraft pilots, SM Thesis, Dept of Aero and Astro, MIT, September 1981
- 7 TOLE J R
STEPHENS A T
HARRIS R L Sr
EPHRATH A Visual scanning behavior and mental workload in aircraft pilots. Aviation Space and Environmental Medicine, January 1982 B
- 8 TOLE J R
et al Visual scanning behavior and pilot workload, NASA CR-3717, August 1983

CHAPTER 10

FLIGHT TEST EVALUATION OF CREW WORKLOAD

by

W A Wainwright
Test Pilots Office
British Aerospace
Hatfield, Herts, UK

PART I

AIRCRAFT CERTIFICATION FOR A MINIMUM CREW OF TWO PILOTS

INTRODUCTION

This paper describes the method developed in 1982 to certificate the BAe 146 for operation by a minimum crew of two pilots to the requirements of JAR 25.1523. The method was based primarily on subjective assessment of workload but employed objective data to support that assessment. All the data were collected from one flying phase and no flight or ground simulator assessments were performed, neither were the results correlated with any previous evaluation. In this respect, the evaluation of the BAe 146 was unique amongst civil workload certification programmes.

The flight evaluation was conducted as a mini-airline exercise similar to that done by other aircraft manufacturers (4) (5). Three teams of two pilots flew consecutive three day intensive flight schedules around a circuit of 3 major European airfields, London — Heathrow, Paris — Charles de Gaulle, Amsterdam — Schiphol, with crew duty hours on some days considerably in excess of those normally allowed for passenger carrying operations. The flight schedules called for multiple legs in a high density air traffic environment, and thus the evaluation was concentrated upon the high workload phases of airline operation. Additionally, operations with inoperative items from the Minimum Equipment List in conjunction with in-flight failure conditions were assessed. Deliberate in-flight failures or dispatch inoperative items were not simulated on the first day of each teams participation in the valuation, but those unplanned events that occurred were logged. Prior to the evaluation, preview flights around the exercise route were arranged for each crew in a BAe 125.

The style of evaluation was evolved from the following considerations:

- 1 The purpose of an assessment of crew workload is to evaluate the workload experienced in flight, and all predicted values of workload obtained from other sources must be confirmed by an in-flight assessment. Therefore all ground-based assessment techniques, including flight simulation, are design tools and not part of the final proof of adequacy, and we resolved to concentrate all our effort into an *In-Flight Evaluation*.
- 2 Any assessment exercise that was not prohibitively expensive was considered to represent such a small sample of experiences that it would only be relevant if the environment was so demanding that it would be conducive to exposing weaknesses in the design. Thus the measurement of workload under normal benign conditions was disregarded and we concentrated on the *high workload regions of aircraft operations*. These were considered to be the arrival and departure phases at very busy airports.
- 3 No single evaluation technique was considered to be sufficiently reliable on its own to be the sole arbiter of acceptability. Thus a matrix was developed incorporating *several individual indicators of workload* that were correlated to give an overall assessment of acceptability.
- 4 There is no totally accepted definition of workload, but studies have shown that most pilots prefer a definition that relates workload to effort (4). Equally there is no generally accepted objective measure of effort, although heart rate monitoring has been successfully used in co-ordination with subjective assessment (5). We therefore chose to base our evaluation on *subjective pilot assessment correlated with heart-rate monitoring* as the principle indicators of workload.
- 5 A high workload was considered to be more of a problem to a tired pilot, and therefore an excessive workload is more likely to be revealed when the crew is fatigued. Thus our participating crews were asked to fly *long duty days with minimum rest periods*.

The decision to use subjective assessment was based on the following assumptions:

Workload is best defined as related to effort.

The most accurate measure of the effort expended by an individual is that individuals subjective assessment.

However, subjective assessment is subject to the accusation of bias since it is produced by individuals who have many different and conflicting interests. Therefore some objective support is required for a subjective evaluation. The technique of using heart-rate to evaluate workload developed by Roscoe at RAE Bedford, England (5) was chosen as the most reliable indicator of effort by an objective technique.

Heart-rate monitoring was used in a supporting role because the idiosyncratic nature of heart-rate and the differing arousal behaviour of individuals mean that heart-rate monitoring cannot be used as an absolute indicator of workload although it can be used to validate subjective opinion.

AD-P005 637

Two basic methods exist to measure workload by subjective assessment — Questionnaires and Rating Scales. Questionnaires can only give a broad assessment which is based more on the feelings after the event, but rating scales can be used to give instantaneous impressions that are not subject to fading with time, and the individual ratings can be assembled to give a detailed record of the flight. Therefore assessment by use of a rating scale was chosen as the primary indicator of workload.

The Roscoe-Ellis rating scale (Figure 1) was derived from the well established Cooper-Harper rating scale (6) to be a scale specifically to rate pilot workload. Its use was pioneered at the Royal Aircraft Establishment where it was used during the Economic Category 3 programme and other flight trials (7) (8). The BAe 146 Workload Evaluation was the first use of the scale on a civil certification programme, and it proved to be eminently suitable for this purpose.

PILOT WORKLOAD RATING SCALE
(for a specified piloting task)

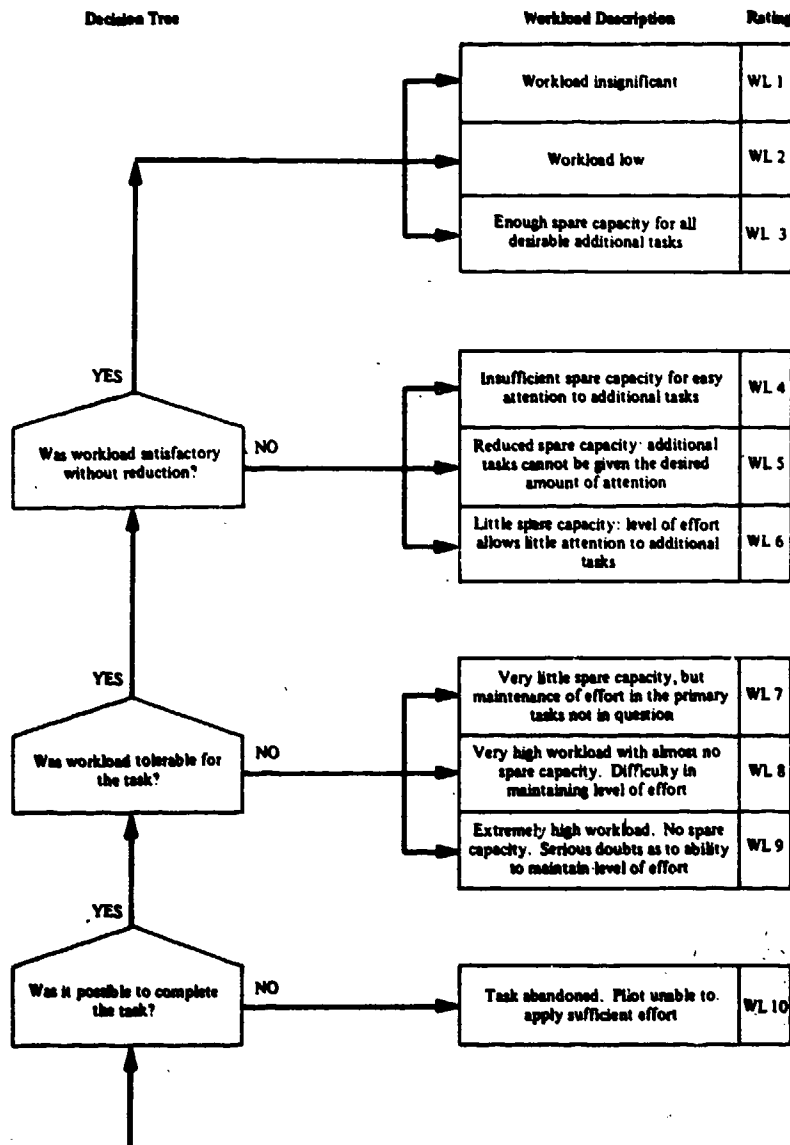


Fig.1 Pilot workload rating scale

The pilot starts his decision-making process at the bottom left corner of the decision tree

METHODOLOGICAL CONSIDERATIONS

Personnel

The following were considered to be essential requirements for crew members involved in a flight test evaluation of workload:

- Experience on the aircraft to be evaluated.
- Experience in evaluation techniques.
- Experience in airline flying.

Where the evaluation includes an element of subjective assessment there exists an additional but paramount requirement for impartiality. Although it is not needed for any scientific reason, the principle objection raised against subjective assessment as a means of evaluating crew workload is the possibility of bias by the subject pilots.

The above requirements are contradictory for an aircraft that is being evaluated during the pre-certification phase of its development because the only pilots with the requisite experience on type will be Company test pilots, who may not be considered to be impartial by external parties. Furthermore there will be very few totally impartial pilots, who have any experience on the aircraft. Thus it is necessary to use a melange of pilots — all of whom satisfy some of the criteria, but none of who satisfy them all — in order to form a balanced evaluation team. A mix of BAe and CAA pilots was used for the evaluation of the BAe 146.

Observers were carried on the flight deck for the following reasons: firstly, to enhance the impartiality of the evaluation, and secondly to broaden the scope of the assessment. Thus the observers had to satisfy the following criteria:

- Independent of any commercial connection with the aircraft manufacturer.
- Expert in either human factors, flight operations or airworthiness.

Assessment of Workload

An acceptable workload was demonstrated by the following means, the results from which were correlated to give an overall view of the workload experience during the evaluation:

- 1 Efficient Operation during an intensive 'mini-airline' schedule by several crews of two pilots.
- 2 Safe Operation in a high density Air Traffic environment by several crews of two pilots.
- 3 The ability of two-pilot crews to operate the aircraft safely and without physical or mental fatigue on demanding duty schedules.
- 4 Subjective pilot opinion of conceived workload
- 5 Subjective opinion of perceived workload by independent aviation experts
- 6 Analysis of pilot heart-rate.

The aircraft was fitted with instrumentation to the scale described in Part II of this chapter.

Assessment of Performance

The aim of any airline is to run an efficient operation and one indicator of efficiency is the ability of the aircraft crew to operate to a demanding schedule. Thus the workload involved in operating the aircraft must be amenable to the achievement of such an aim. Therefore the ability of the crew to keep to the schedule is an indicator of an acceptable workload, although one that would be unconvincing if required to stand on its own. The schedule established for the mini-airline exercise was demanding (7 legs on each of the first 2 days with a 13½ hour duty day, and 4 legs in the third day with a 7½ hour duty day), and all deviations from it were logged. The deviations were correlated with other data to establish if any could be attributed to an excessive crew workload.

Another aim of any airline is to run a safe operation. The achievement of this can be assessed by the occurrence or otherwise of ATC violations or potentially dangerous incidents. The observers were tasked with logging any incidents or violations and any crew errors and deviations that could have led to incidents or violations under other circumstances. All occurrences of errors were correlated with other data to establish if any could be attributed to an excessive crew workload.

The efficient and safe operation of the aircraft by crews of 2 pilots was achieved when operating to duty hours well in excess of those normally permitted for public transport crews. A lack of crew errors was considered to be indicative that no undue mental stress was experienced, and was confirmed by the analysis of pilot heart-rate which also indicates whether any physical fatigue occurred.

The principle measure of conceived workload was the *pilot rating scale* and the use of it is described in more detail in Part II. Other subjective data were obtained from questionnaires. A post-flight questionnaire was completed after every sector by each pilot. This questionnaire asked for information about the ATC and weather situation, about the level of workload experienced in each phase of flight, and asked for an opinion on the cause of any high workload that was experienced — the available causes ranged from difficult ATC environment to poor aircraft handling. An opinion was also demanded of the workload experienced compared to that previously experienced on a similar type of aircraft under similar conditions. Finally post-exercise questionnaires were completed by each pilot which called for a more general appreciation of workload attributed to individual features of the aircraft. The specific criteria of JAR 25 Appendix D were used as a basis for this questionnaire.

The observers were used as a source of perceived workload. They gave rating scores for each pilot using the pilot rating scale, completed a post-flight questionnaire that was similar to that of the pilots, and completed a post-exercise deposition that gave their general impression of the workload experienced by the pilots.

Heart-rate was continuously recorded for each pilot from starting engines to shutting down at the end of each sector. The rating scores given by pilots and observers were superimposed on the heart-rate trace together with a time base so that the results could be correlated. Mean heart-rates for the 30 seconds preceding rating scores were compared to the ratings, and the instantaneous heart-rates were examined for evidence of rapid variations in rate that could suggest sudden changes in workload. Heart-rates are essentially idiosyncratic and it is not possible to compare the heart-rate of one pilot to that of another. It is normal, in fact it is desirable, for heart-rate to rise as a pilot becomes aroused in preparation for the take-off and the landing. Equally, individuals have different arousal patterns in response to changing events. Thus simple comparisons of instantaneous heart-rates are misleading and each pilot has to act as his own control, and data obtained from the preview flights in the BAe 125 around the exercise routes were used to give a base-line behaviour pattern.

RESULTS

The core of the results was the pilot and observer rating scores (see Part 2 of this chapter for details of the analysis, and Figures 3, 4 and 5 for examples).

The results obtained from the analysis of the rating scores were then compared with the following other sources:

- 1 Questionnaires — broad statements on workload levels were obtained from post-flight and post-exercise questionnaires. These statements were compared to the conclusions obtained from the individuals rating scales.
- 2 Error counts — Observers were asked to record any crew errors that they noticed, and the ratings given at the time of the occurrence of an error were examined to identify whether the error could be attributed to a high workload.
- 3 Depositions — Observers completed Depositions on their opinion of the crew workload after they had concluded their participation in the exercise. The opinion expressed was compared to the ratings given by the observer.
- 4 Heart-rate — Rating scores were compared to the heart-rate behaviour appertaining at the same time.

The assessment of workload achieved by the above comparisons was then examined in relation to the following data to establish whether it was consistent with it:

- a. Efficiency of Operations — obtained from analysis of the programme achieved.
- b. Safety of Operation — obtained from the Error Count and Observer Comment.
- c. Physical or Mental Fatigue obtained by analysis of heart-rate traces.
- d. Comparison with Similar types — obtained by questionnaire.
- e. Specific Compliance with JAR 25 Appendix A — obtained by questionnaire.

Finally, the video record was available for examination to resolve any inconsistencies in the results.

CONCLUSION

The flight test evaluation of the BAe 146 used a variety of assessment methods — including practical demonstration, qualitative and quantitative subjective evaluation, subjective comparison with similar aircraft types, and objective physiological evaluation — and all confirmed that the crew workload on the BAe 146 was compatible with operation by a minimum crew of 2 pilots. This result has since been further confirmed by in-service experience and by the subjective assessment of line pilots flying the type in airline service which has been obtained by questionnaire.

No inconsistencies or ambiguities occurred during the evaluation or have appeared since it was conducted. This supports the contention that a single phase flight test evaluation of workload is an adequate and appropriate test of the suitability of an aircraft for operation by its minimum crew. Furthermore the agreement achieved during the evaluation of the BAe 146 between all the types of data collected endorses the integrity of the method of subjective assessment supported by heart-rate analysis.

PART 2

SUBJECTIVE ASSESSMENT CORRELATED WITH HEART-RATE

INTRODUCTION

The technique suggested in this paper was developed for the flight Test Evaluation of workload in the BAe 146 in 1982, described in Part I of this Chapter. This method, of using subjective assessment correlated with heart-rate monitoring, was used as the basis of the evaluation which was then supported by other subjective and objective data to obtain CAA and FAA certification of the aircraft for a minimum crew of two pilots (9).

It is essential that any assessment techniques does not superimpose any extra workload on top of that being evaluated. This aim may be achieved by the following means:

- 1 A simple scoring pad with 10 buttons — one for each rating — should be provided for each individual in a position where it is easy to use and does not intrude into normal operations, for example, situated on the central boss of the control wheel at each pilot's station.
- 2 No mental effort should be required of the participants to remember when to give a rating — they should merely have to respond to a call by an Exercise Controller, who calls for ratings in response to a specific programme.
- 3 The participants must be given practice in using the rating scale prior to the exercise.
- 4 Differentiating between individual ratings has to be easy and consistent. The rating scale (Figure 1) is based on the concept of spare capacity and differentiating between ratings is done by assessing the amount of spare capacity available. Primary and secondary tasks are defined, and thus participants only have to assess how much spare capacity they have available to devote to the secondary task whilst performing the primary task.

The ratings given must be discrete so that an individual is not influenced by his partner. This can be achieved by use of individual scoring pads.

GUIDELINES FOR USING THE RATING SCALE

Instantaneous ratings of workload can be misleading — they can be given during a temporary lull or peak. Therefore, participants should be asked to maintain a continuous estimation of their workload and to give a scale number for the workload pertaining during the previous 30 seconds or so when responding to a call for a rating.

The primary task for each pilot is all those taskloads that are essential to operating the aircraft in his crew capacity of captain or first officer. Thus, the captain's primary task includes all necessary actions to control the aircraft flight path, to manage the flight, and to comply with ATC requirements, including visually searching for aircraft reported by ATC. Similarly, the first officer's primary task includes all necessary actions to operate the aircraft systems, to navigate the aircraft, to manage the radio, and to search for aircraft reported by ATC.

Spare capacity is estimated by assessing the amount of time available for, and the ease of performing, all secondary tasks. Secondary tasks are defined as all those tasks normally performed on aircraft flight decks that are not considered to be part of the primary task. This includes such non-essential duties as monitoring the other pilot and visual lookout not in response to ATC directions, which are actions normally carried out by a competent crew member, but which are often the first duties to be shed when workload is increased.

METHODOLOGY

The purpose of this paper is to describe the use of this technique to evaluate workload during a specific phase of flight — the approach to landing. It is assumed that the aircraft is a passenger carrying airliner configured for a crew of 2 pilots but possessing a seat on the flight deck for a 3rd crew member. The task being assessed is the ability to fly a standard ILS approach to a full stop landing.

The operating crew of two pilots and a suitably qualified observer occupying the 3rd crew seat form the evaluating team. An Exercise Controller co-ordinates the exercise, request ratings according to a pre-determined rating plan (an example of such a plan is shown in Figure 2) and imposes systems failures when demanded by the evaluation programme. The following instrumentation would be appropriate to an evaluation of this kind.

- 1 A data display and storage system should be installed to record the rating scores and heart rate. Instantaneous display of data should be available, together with hard-copy print out.
- 2 A two camera video system should be installed — one camera to view the main panel instruments and centre console and the other to view the overhead panel. A split picture giving views from both cameras should be available. The display format should be controlled from the exercise control position. A video recording should be available for examination post-flight to resolve any confusions in the data.
- 3 An exercise control position should be installed in the cabin of the aircraft containing a video display of the cockpit, a console for the video system, and the rating score system. A communications point should be provided.
- 4 Rating score pads must be provided for each pilot. A suitable position for them would be the centre of control wheel. Individual push-button electros must be available for each rating number (1-10) and a cancel button must be provided. A light on each pad should illuminate when a rating has been requested and extinguish when the rating is given (the primary

ratings will be requested at the following events:

1. Intercepting the glidepath
2. Overhead the Outer Marker
3. 1000 ft above touchdown
4. 400 ft above touchdown
5. On the runway - decelerating through 80 kts
6. Turning off the runway

Fig.2 Workload rating plan for evaluation of workload on final approach

means of calling for a rating is by verbal command over the aircraft intercommunications system). The observer should be provided with two score pads — one for each pilot.

5 Plug points must be provided at pilot's seat to transmit heart-rate signals to the data system.

The subject pilots fly the aircraft in accordance with standard operating procedures maintaining a constant awareness of their spare capacity in accordance with the established guidelines. The observer monitors their actions and maintains his own assessment of their workload. At the appropriate point on the flight path, the Exercise Controller requests a rating. The pilots and the observer give their response on their score pad. The Exercise Controller confirms that 4 score have been received. The lights on the score pads are available to the pilots and the observer for them to confirm if they have responded. No other tasks are required of the evaluating team.

The data — heart-rate and rating scores — are recorded automatically for future analysis.

INTERPRETATION OF RESULTS

The first analysis of the results should include a correlation of all ratings that stand out from the general scale with other available data. This would include subjective data from other sources and objective data such as the heart-rate pertaining at the time. A comparison between the pilot's own rating and the observer's rating for him is particularly useful, and the time taken to respond to the call for a rating should be considered. This first analysis should eliminate any ratings that are obvious selection errors and could attribute instances of high workload to particular problems such as weather, ATC, or simulated system failures.

A second and simple use of the ratings is to merely examine a graphical plot and note the predominance of particular scores. But in this context, it is important to consider the distinction between satisfactory and acceptable on the rating scale. The scale was primarily designed for use in aviation flight research, although the concept of spare capacity appears to be ideal for workload certification. However, the workload description for a score of 4 describes a situation that is satisfactory for multi-crew aircraft operating in a high workload environment where it is normal practice to allocate priorities and to time-share between tasks. Thus, a satisfactory workload is not only demonstrated by all ratings falling in the range 1 to 3, but also when the mean workload is in that range, with some deviations into the acceptable bracket.

A more sophisticated use of the ratings is to compare homogenous blocks of aggregate ratings and histograms can be used for this. The following comparisons are suggested:

- 1 Comparison of a pilot's ratings with those given for him by the observer. (An example is shown in Figure 3).
- 2 Comparison of individual pilot's ratings. This, combined with 1, can be used to prove a lack of bias within the results or to correct for it. (An example is shown in Figure 4).
- 3 Comparison of a pilot's ratings for normal flights with those for flights when simulated failures are included.
- 4 Comparison of ratings for Captains with those for First Officers to establish the balance of workload within the flight deck. (An example is shown in Figure 5).

A statistical analysis of the ratings involving the calculation of a mean and standard deviation for each subject is considered to be inappropriate because it could reproduce a mean rating that is not an integer and a deviation that is a fraction of a rating score. Both are meaningless in the context of the rating scale, and it is considered that an overall view of the accumulated ratings obtained from graphical representation or by means of histograms is the only valid interpretation of the results.

The final correlation of the ratings is with heart-rate (10). The following observations on the use of heart-rate data in the assessment of pilot workload are based on the results of the BAe 146 evaluation:

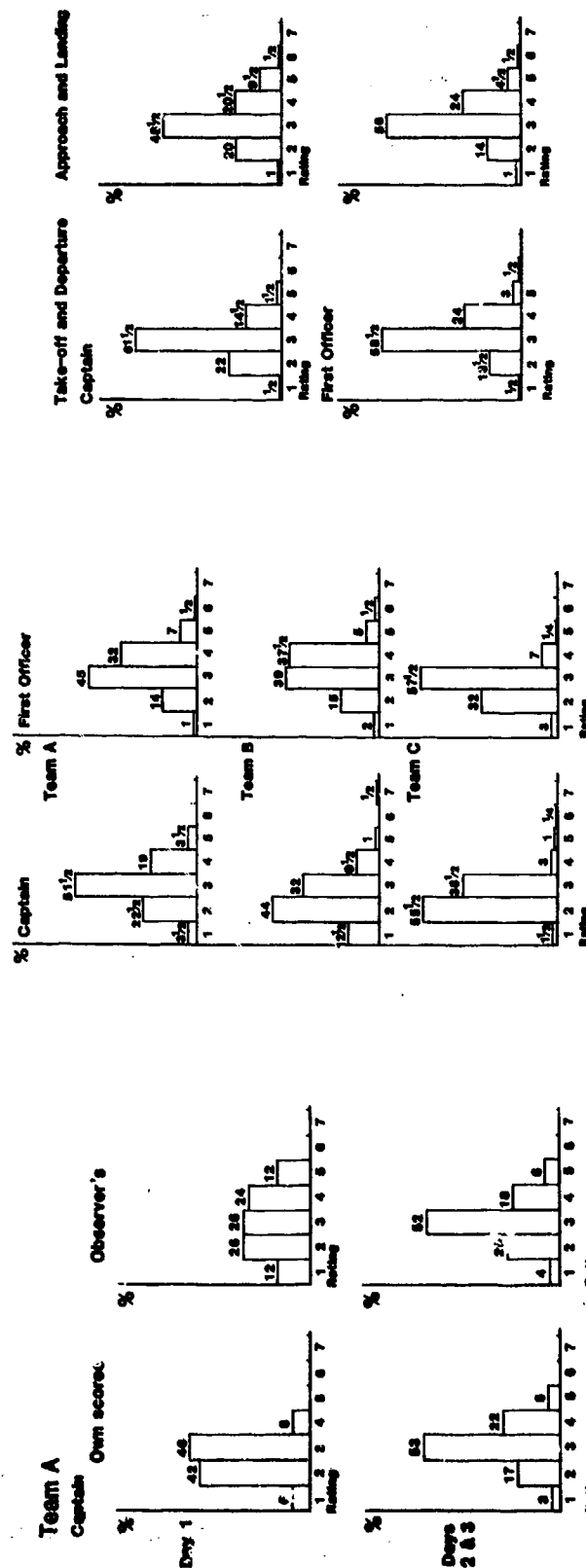


Fig.3 Example of histograms of rating scores

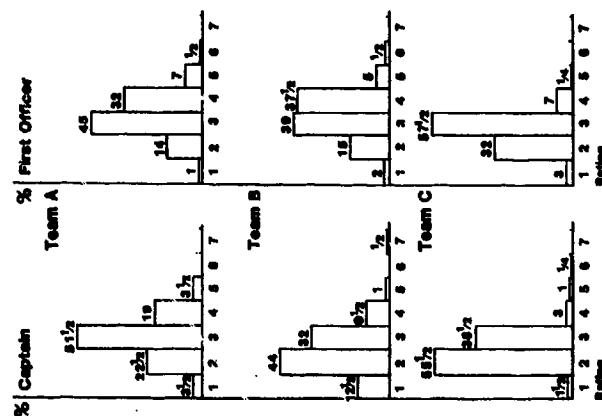


Fig.4 Histograms of rating scores: comparison for impartiality

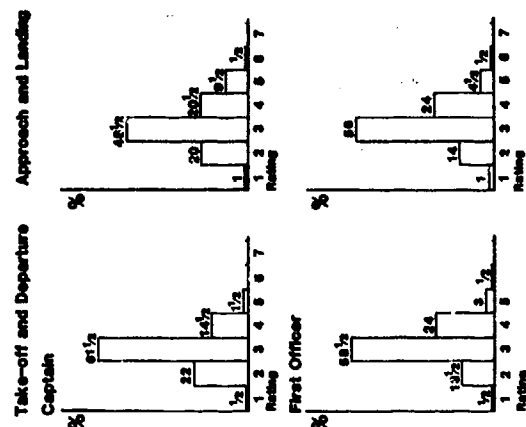


Fig.5 Histograms of ratings for high workload phases of flight

- 1 Pilot's heart-rate levels tend to augment their ratings of workload; in particular, there should be no significant disagreements between the two. Some minor differences may occur when ratings are given without considering the workload during the preceding 30 secs.
- 2 Examination of the heart-rate plots should reveal evidence of unduly high levels of workload. An overall increase in heart-rate would suggest stress from time pressure or pacing whereas a decrease in the heart-rate variability (sinus arrhythmia) would indicate increased mental activity.
- 3 A comparison of heart-rate levels recorded at various times during the working day would indicate whether workload was influenced by fatigue.

PROBLEMS IN USING SUBJECTIVE RATING SCALES

It is worth noting the following problems which were experienced during the application of this technique in the evaluation of the BAe 146.

Subjective rating scores are very idiosyncratic. We evolved criteria for the use of the rating scale in an attempt to remove individual variations in approach to the rating — but not everyone follows the criteria. For example one observer commented that the rating system did not reflect task sharing, indicating that he had not understood the criteria that asked for workload to be related to the previous 30 seconds.

One would expect the observers' scores to be generally lower than the pilots' own scores — since mental workload will not be apparent to an observer. However, the above relationship was completely reversed for two pilots, whilst one pilot scored a mean almost identical to the observer. One would also expect the workload to increase with the difficulty of the task ie when coping with in-flight failures and dispatch with inoperative equipment. But, in the event, no change occurred for five pilot's. This can be explained by increased familiarity with the ATC environment counter balancing increased difficulty in the task. However one pilot's ratings actually reduced as the exercise progressed and one must then suspect a change in his own datum albeit only by one point in nine.

Unfamiliarity of the pilot with the aircraft can affect the results. One observer commented that one incident where a First Officer was experiencing a high workload was obviously due to the pilot's unfamiliarity with the aircraft. Overlaid over everything else, there always exists a difficulty in being consistent when asked for instantaneous decisions, and this is bound to introduce some scatter in the results. And there is the problem of exercise artificially. How do you rate workload for an emergency situation that would normally demand an emergency call and special treatment from ATC when the pilot has to handle the simulated emergency and conform to normal ATC? But the main problem is that of bias in the subjects. We used impartial pilots as a form of bias control and we obtained reasonable correlation between their scores and the scores of the BAe pilots.

Finally, these problems reflect the difficulty in using subjective assessment as an absolute measure. However, with meticulous cross-references between the rating scales and correlation with other data such as heart-rate, a valid overall picture of the workload experienced in flight can be drawn. But the problems will be exacerbated if a discontinuity is introduced as will occur if an attempt is made to compare different aircraft or different environments without the teams of participants remaining identical throughout all phases of the evaluation.

REFERENCES

- 1 SULZER R L Flight crewmember workload evaluation. US Department of Transportation, FAA-RD-129, 1981
COX W J
MOHLER S R
- 2 McLUCAS J L Report on the Presidents Task Force on aircraft crew complement, 1981
DRINKWATER F J
LEAF H W
- 3 SPEYER J J Dynamic workload analysis flight campaign. A300FF Certification Flight Test Report, Airbus Industrie A1/V-F 1306/81, 1981
FORT A P
- 4 ELLIS G A The airline pilot's view of flight deck workload: A preliminary study using a questionnaire. RAE Technical memorandum FS(B)465, 1982
ROSCOE A H
- 5 ROSCOE A H Assessing pilot workload in flight. In: Conference Proceedings No 373 AGARD, Paris, 1984
- 6 COOPER G E The use of pilot rating in the evaluation of aircraft handling qualities. NASA Report TN D-5153, 1969
HARPER R P
- 7 LUMSDEN R B The economic Category 3 Programme 1975-80. RAE Technical Report 81025, 1981
- 8 H Heart Rate as an in-flight measure of workload. In Proceedings of USAF AFFTC/NASA DRYDEN/AIAA Workshop on Flight Testing to Identify Pilot Workload and Pilot Dynamics, Edwards AFB California, 1982

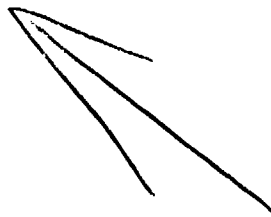
68

9 WAINWRIGHT W A

BAe 146 — Flight test evaluation of workload. Certification Report British Aerospace
HTD R 460-00 SC0038, 1983

10 ROSCOE A H

Analysis of heart-rate data. In: Certification Report British Aerospace HTD R 460-00
SC0038 Annex K, 1983



CHAPTER 11

MEASUREMENT OF AIRCREW WORKLOAD DURING LOW-LEVEL FLIGHT

by

I Gavin Lidderdale*
Principal Psychologist (Human Factors).
Headquarters Strike Command
Royal Air Force

PART I

A COMPARISON BETWEEN IN-FLIGHT AND POST FLIGHT ASSESSMENT METHODS

INTRODUCTION

The operational evaluation of modern military combat aircraft requires aircrew to operate and monitor complex systems whilst flying at ultra low-level, often at night in poor weather and in hilly terrain.

In this demanding environment, the development of crew cooperation procedures and the integration of these with new tactical manoeuvres has pushed aircrew to the limit of human performance. The continuing development of sensors, weapons and flight control systems will place even higher demands on aircrew in the coming years and it is for this reason that the development of practical and reliable methods for monitoring cockpit workload is a high priority.

LOW LEVEL FLIGHT TRIALS

During recent flight trials of an advanced combat aircraft with a two-man crew, a combination of heart rate recordings and in-flight workload ratings provided measurements of workload during critical low level manoeuvres. The physiological recordings provided detailed continuous data which were useful in identifying short duration increases in workload and also gave some indication of the differences between mental concentration and psychomotor activity (1). The subjective assessments were made for critical flight manoeuvres and represented a summary of the workload for a particular phase of flight. As such these assessments proved to be useful in identifying the absolute, perceived, workload and for making comparisons between different aircrew performing the same task.

However, as flights were made at lower flight levels and over difficult terrain at night or in poor weather, it was thought that the aircrew would not be able to make in-flight workload assessments. In anticipation of this potential problem a method was devised for obtaining accurate workload assessments during the post-flight debrief.

Despite initial doubts, the aircrew were able to give in-flight assessments even in the most demanding environments. As a result, the technique developed for these low level trials provided data from both in-flight and post-flight assessments. Although there are many recommended techniques for assessing workload, there is little published data from flight trials and even less which compares the different methods (2). The trials described below created the opportunity for such a direct comparison.

IN-FLIGHT WORKLOAD ASSESSMENT

An aircraft mission profile was developed which continued 10 elements of low-level flight which were of operational interest. For the sake of simplicity, these tasks will be referred to as elements A to J. The trials were designed to assess workload and develop crew cooperation procedures during terrain following flight in a modern two-man combat aircraft.

The method chosen for the in-flight assessment of workload was an adaptation of the Cooper-Harper rating scale (3). The method was pioneered by Roscoe and Ellis (4) (5) at the Royal Aircraft Establishment at Bedford and is sometimes referred to as the 'Bedford Scale'. The technique relies on aircrew making a subjective judgement about their workload using the hierarchical decision tree and rating scale shown below. It was found that aircrew understood the scale readily and whilst it was sufficiently comprehensive to cover all circumstances it was easy to remember and small enough to be carried on the flying suit kneepad. (See Figure 1, Chapter 10.)

Before each sortie, the aircrew were rehearsed in the definitions of the scale so that it would be more easily recalled during the flight. In addition, a copy of the scale was attached to the flying suit kneepad so that it could be referred to if necessary.

Independent assessments An important requirement of the trials was to obtain assessments which were, as far as possible, independent. In order to achieve this, the navigator was required to record his own assessment before asking the pilot to rate the workload in the front cockpit. It was found that the navigator was always able to perform this task without degrading his primary duties whereas the pilot was often unable to write notes.

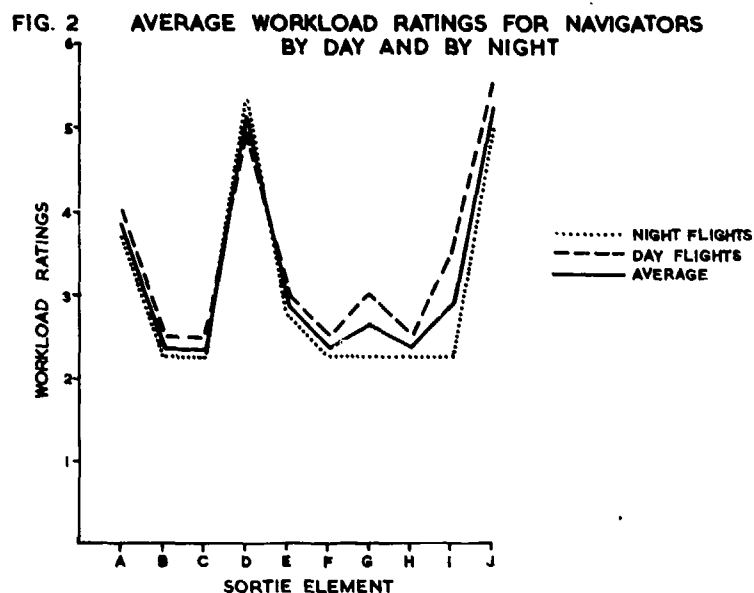
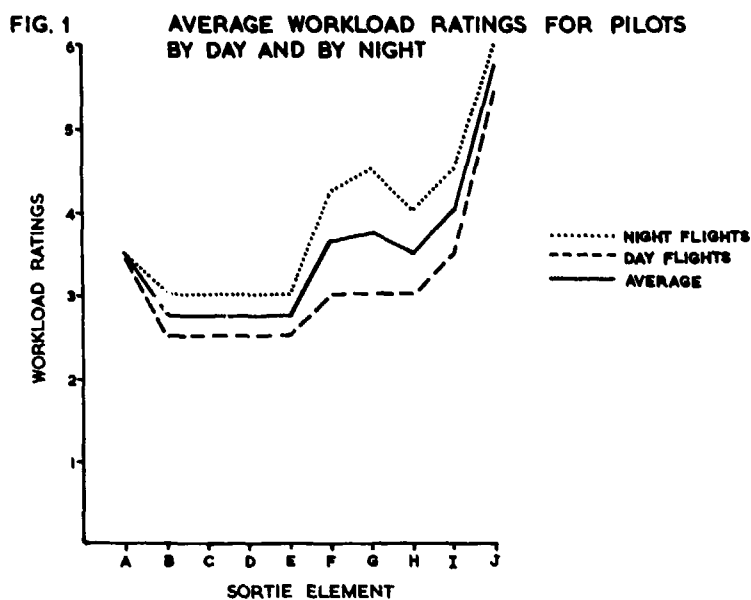
A cockpit voice recorder was also used and this provided additional and useful information. The tapes often provided the key to otherwise unexplained rises in heart rate and also acted as a record of crew commentary which was useful in developing more effective cooperation procedures.

* Present address: ACCIS Studies Branch Information Systems Division, Supreme Headquarters Allied Powers Europe Technical Centre, PO Box 174, The Hague, The Netherlands

AD-P005638

Results of in-flight workload assessments Two crews, each comprising a pilot and a navigator, took part in the flight trials. Each crew flew a specified route both in day time and at night on separate days. The route was designed to contain 10 key mission elements, A to J, which were of operational interest and each element was assessed using the 'Bedford Scale' described above.

The results of the trials are summarised in graphical form as Figures 1 and 2 below.



POST-FLIGHT WORKLOAD ASSESSMENT

The use of assessment techniques which require aircrew to make judgements during flight necessarily means that some attention is diverted from the primary task, even though this may be for a very small period of time. This concern stimulated the search for a post-flight workload assessment technique which would give equivalent ratings for the phases of flight under consideration.

The 'Bedford Scale' (4) (5), which had proved to be useful for in-flight assessments was found to be an inappropriate tool for use during the post-flight debrief. The primary reason for this stems from the finding that when ratings are made after flight, aircrew find great difficulty in reconstructing the complex events of each flight element in sufficient detail to be certain of their response. What is possible, however, is for aircrew to make a relative workload comparison between any two elements. Based on this finding, a method was devised which reduced the assessment task to the level of pairwise comparisons and yet which enabled the investigator to reconstruct the sortie workload from these results.

The method chosen for this task was based on the Analytical Hierarch Process reported by Saaty (6). This method is used to analyse pairwise comparisons made from subjective ratings and avoids the problems associated with absolute rating scales which have limited use in a post-flight context.

The trial consisted of 10 flight elements which were to be assessed. Taking each possible pair of elements in turn as described below, the aircrew were asked to assess which was higher in workload and, unless they were equivalent, by how much. A five point scale was chosen to describe the relative workload as shown in Table 1 below.

TABLE 1 RELATIVE WORKLOAD ASSESSMENT SCALE

- 1 EQUAL WORKLOAD
- 2 SLIGHTLY HIGHER WORKLOAD
- 3 MODERATELY HIGHER WORKLOAD
- 4 VERY MUCH HIGHER WORKLOAD
- 5 EXTREMELY HIGH RELATIVE WORKLOAD

The Saaty Method As the trial sortie contained 10 elements which were to be assessed, a clear and concise method of presenting all possible combinations of pairs of elements is as a matrix; in this case a 10 x 10 matrix. The matrix is symmetrical about the diagonal line as shown below with 45 unique combinations in the half matrix.

(It should be noted that when 'n' alternatives are to be compared, the number of comparisons is $\frac{1}{2}n(n-1)$. Although each assessment is an easy procedure, when there are a large number of elements the number of pairs can be very large.)

EXAMPLE OF THE POST-FLIGHT DEBRIEFING METHOD OF WORKLOAD ASSESSMENT

The following example of the use the post-flight debriefing method for workload assessment is given as an illustration of how the technique can be applied in practice.

For a matrix of 4 elements, as shown in Table 2 below, each pair in the bottom half of the matrix is assessed using the relative workload scale in Table 1 above. The element on the vertical scale is compared with the element on the horizontal scale: if the element on the vertical scale is the higher in workload, a positive number is entered. When the element on the vertical scale is lower in workload the number entered is the reciprocal.

TABLE 2 EXAMPLE OF A 4 X 4 MATRIX AS COMPLETED DURING POST-FLIGHT DEBRIEF

		SORTIE ELEMENT			
		A	B	C	D
SORTIE ELEMENT	A	1			
	B	2	1		
	C	3	1/2	1	
	D	2	1	4	1

In the example in Table 2, element 'B' is rated as a '2' compared to element 'A'; this means that B had a slightly higher workload than A. The comparison between elements 'C' and 'B' resulted in an assessment of $\frac{1}{2}$ which indicates that C had a slightly lower workload than B.

Once the aircrew had completed the lower triangular matrix, it is a simple procedure to fill in the upper triangular matrix as the reciprocal values. The matrix at Table 2 then becomes the completed matrix as shown in Table 3.

TABLE 3 EXAMPLE OF 4 X 4 MATRIX AFTER COMPLETION BY ANALYST

		SORTIE			
		A	B	C	D
SORTIE ELEMENT	A	1	1/2	1/3	1/2
	B	2	1	2	1
	C	3	1/2	1	1/4
	D	2	2	4	1

The analysis of the matrix would normally require complex algebra which is better undertaken using a micro-computer program. The technique is described by Saaty (6) and a simple computer program is reported by this author (7). However, a simpler, and broadly comparable result can be obtained using the following method:

- a. Given the matrix of 4 x 4 elements as shown above, the aircrew completes the lower triangular matrix to give the results

as shown in Table 3 and the analyst then completes the matrix so that the upper triangular matrix is the reciprocal of the lower half. The result is now as shown in Table 3.

- b. From the completed matrix, product of the numbers in each row is calculated.
- c. The n th root of the product is computed where n is the number of elements in the matrix; in this example the 4th root is taken.
- d. The roots are then summed to give a total.
- e. Each product root, from c, is then divided by the summed total of product roots to give a weighted product. The sum of all weighted products will then sum to unity.

A worked example from the matrix at Table 3 is shown below.

TABLE 4 EXAMPLE OF ANALYSIS OF 4 X 4 MATRIX

SORTIE ELEMENT	SORTIE ELEMENT				PRODUCT ROOT WEIGHT		
	A	B	C	D			
A	1	1/2	1/3	1/2	0.008	0.3	0.72
B	2	1	2	1	4	1.414	0.338
C	3	1/2	1	1/4	0.375	0.782	0.187
D	2	1	4	1	8	1.682	0.403
Sum of roots =					4.1778	1.000	

The results from this method give a comparable solution to the more complex procedures described by Saaty (6) and this has the advantage of rapid analysis and feedback of results.

CONSENSUS METHOD OF COMBINING WORKLOAD RATINGS FROM THE POST-FLIGHT METHOD

When it is necessary to combine the results of two or more subjects, the weightings can be averaged using the consensus method described below.

As an example of the consensus method, the weights of five subjects is taken for 4 sortie elements A to D. Having calculated the weights as described above, the results are placed in a table as shown in Table 5.

TABLE 5 CONSENSUS METHOD FOR COMBINING WORKLOAD WEIGHTS

	SUBJECTS					
	1	2	3	4	5	
SORTIE ELEMENTS	A	0.58	0.29	0.28	0.23	0.58
	B	0.09	0.10	0.05	0.07	0.04
	C	0.29	0.56	0.60	0.63	0.29
	D	0.04	0.05	0.07	0.07	0.09
COLUMN TOTALS		1.00	1.00	1.00	1.00	1.00

Having completed the table as shown above, the rows are now rearranged in ascending order of weights to produce a new table as shown below at Table 6.

TABLE 6 CONSENSUS METHOD OF COMBINING WORKLOAD WEIGHTS

SORTIE ELEMENTS	0.23	0.28	0.29	0.58	0.58
	0.04	0.05	0.07	0.09	0.10
	0.29	0.29	0.56	0.60	0.63
	0.04	0.05	0.07	0.07	0.09
COLUMN TOTALS	0.60	0.67	0.99	1.34	1.40

Now, by interpolation between the 2 columns whose totals are astride unity, a new set of weights is computed which sum to unity. In the above example, columns 3 and 4 straddle unity and interpolated weights which sum to unity are computed as below:

A	0.30
B	0.07
C	0.56
D	0.07
TOTAL	1.00

Results of the post-flight workload assessments The results of the post-flight assessments of relative workload using the method described above are summarised in graphical form in Figures 3 and 4 below.

FIG. 3
CONSENSUS WEIGHTINGS FOR PILOTS
BY DAY AND NIGHT

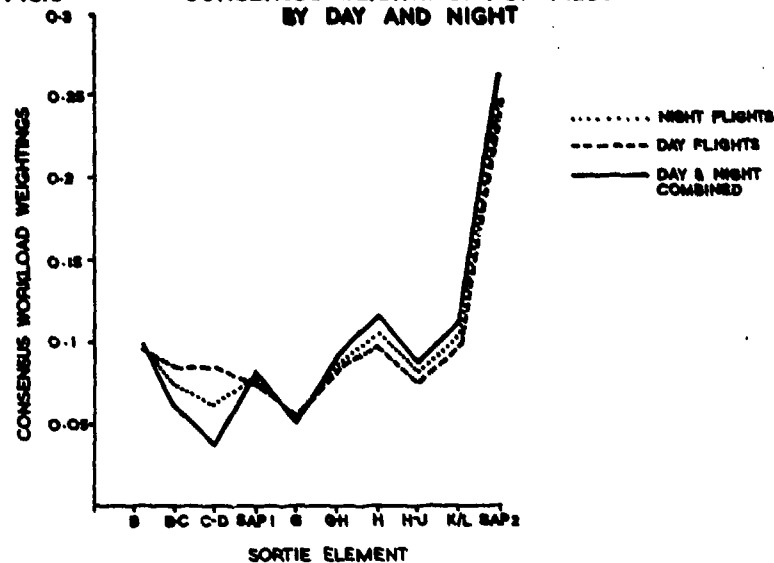
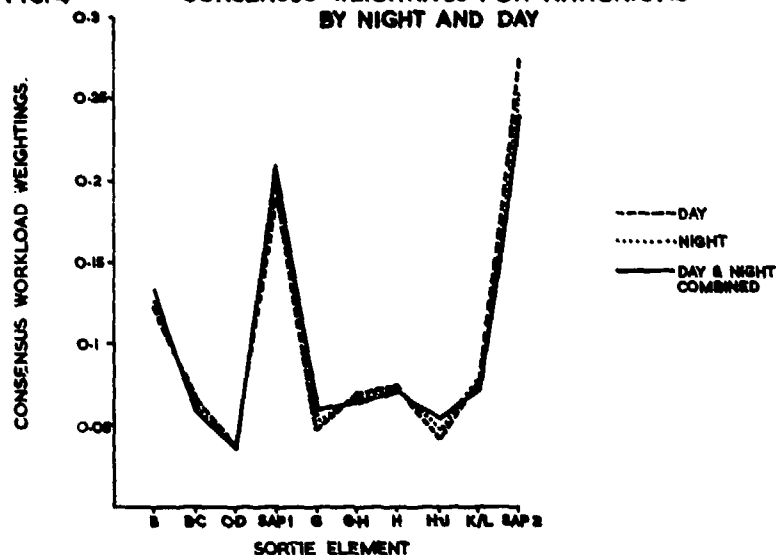


FIG. 4
CONSENSUS WEIGHTINGS FOR NAVIGATORS
BY NIGHT AND DAY



COMPARISON BETWEEN THE IN-FLIGHT AND POST-FLIGHT METHODS OF WORKLOAD ASSESSMENTS

By comparing the results shown in Figures 1 with Figure 3 and the results in Figure 2 with Figure 4 the similarity between the methods is readily apparent*. Clearly, the results can never be the same as, while the in-flight rating scale gives an absolute measure of workload. The post flight method is a relative assessment. However, the similarity of the results tend to suggest that the modified Cooper-Harper or 'Bedford Scale' may also produce results which are relative to some extent. It is possible that all assessments of workload are made from a baseline of comparisons with other elements in the flight and, if this is the case, all rating methods may be relative.

* Using a Spearman Rank Order correlation, the coefficient for pilots was 0.86 and the coefficient for navigators was 0.85; both significant at the 0.01 level.

SUMMARY AND CONCLUSIONS

The paper reports the results of a comparative study of the use of in-flight and post-flight methods of subjective workload assessments in a modern military combat aircraft. The assessments were made during a demanding low-level flight task which was undertaken to assess workload and define crew cooperation procedures for pilots and navigators during terrain following flight.

The in-flight workload assessments were made using a modified version of the Cooper-Harper scale which is referred to as the 'Bedford Scale' (4) (5). Post-flight ratings were made using a method of pairwise comparisons based on a method reported by Saaty (5).

Other measure, including physiological recordings and voice tapes were also taken during the trials to provide additional data.

From the results of the trials, it was found that both methods of subjective workload assessment produced similar results and a rank-order analysis gave high correlations.

The use of the 'Bedford Scale' was found to be easily understood and gained ready acceptance from the aircrew on the trial. Despite initial doubts, it was found that both pilots and navigators were able to give assessments of workload during flight, based on the scale, even under the most demanding flight conditions. By adopting a technique whereby the pilot passed his rating to the navigator over the intercom, the assessment task did not interfere significantly with the primary task. In single seat operations, however, the use of a rating scale during flight may be more problematical because although a voice recorder could be used to record the pilot's rating, there would not be a second crew member to ask for an assessment — an essential requirement during the trials reported in this paper.

Using an adaptation of the pairwise method of comparisons reported by Saaty, a post flight workload rating system was devised which was found to be easy and quick to administer. An analysis technique was also devised which produced results without the need for sophisticated computing power. This technique was found to give relative workload assessments which could be compared to the in-flight ratings and it is proposed that this method could be used in circumstances where in-flight assessments are not possible.

The results reported in the flight trials described above are based on limited data and it would be unwise to draw firm conclusions from this initial evidence. However, there seems to be sufficient cause for continuing with both methods of assessment as they appear to produce comparable data and give greater flexibility to the research scientist. The 'Bedford Scale' has proved to be a practical solution to in-flight workload measurement even during the most demanding tasks. However, for single-seat operations the post-flight method may prove to be the only alternative.

Note: The author wishes to acknowledge the advice and guidance of Dr Alvin Roscoe of Britannia Airways and Peter Haysman of the Royal Ordnance Future Systems Group in the development of the methods reported in this paper.

REFERENCES

- 1 ROSCOE A H Handling qualities, workload and heart rate. AGARDograph No 246 — Survey of methods to assess workload
- 2 WIERWILLE WW
WILLIGES R C
SCHIFFLETT S G Aircrew workload assessment techniques. AGARDograph No 246 — Survey of methods to assess Workload
- 3 COOPER G
HARPER R The use of pilot rating in the evaluation of handling qualities. NASA Tech Note TNXD-5753, Washington DC; 1969
- 4 ELLIS G A Subjective assessment pilot opinion measures. A H Roscoe — Ed AGARDograph AG233 — Assessing Pilot Workload
- 5 ROSCOE A H Assessing pilot workload in flight. Conference proceedings No 373 — AGARD 1984
- 6 SAATY T L The analytical hierarchy process. McGraw-Hill 1980
- 7 LIDDERDALE I G
KING A H Analysis of subjective ratings using the analytical hierarchy process; a micro computer program. OR Branch NFR 1985, HQ STC, RAF High Wycombe

PART 2

Fi P-69 APPLICATION OF THE WORKLOAD MEASUREMENT TECHNIQUES TO A RECCE/ATTACK TASK FOR FAST JET AIRCRAFT (SINGLE PILOT)

A hypothetical Recce/Attack task for fast jet aircraft has been chosen to illustrate the application of the workload measurement techniques described in Part I of this chapter.

BACKGROUND

The choice of workload measurement technique depends both on the task and on the purpose for which the task is being assessed.

- 1 *Task constraints* For single seat operations, the main constraints will be those of limited space (for stowing recording equipment) and the unavailability of prompting and note taking by another crew member.
- 2 *Study requirements* The purpose for which workload is being assessed will have an overriding impact on the choice of measurement techniques and on the way in which the task is subdivided into elements. As an example of a suggested method, it will be assumed that an investigation is being conducted into the task loading of the operational procedures of a single seat pilot during the ingress, weapon release and egress from a target.

WORKLOAD MEASUREMENT METHODS

The Recce/Attack task in a single seat aircraft is a high workload procedure which contains many individual elements. Some of the elements are very short in duration and are preceded and/or followed by other high workload elements. For this type of task, it would not be possible or desirable to require subjective ratings for all elements. Equally, physiological recordings in isolation from such ratings are difficult to interpret. A combination of techniques seems to offer the best solution in this case and the following technique is proposed:

1 *Physiological measurement*

Heart rate recording through the task would give a continuous record of events and provide data for all individual items of the task. The size constraints of the cockpit and the flying clothing worn by the pilot would constrain the choice of equipment although there are small recorders available, such as the Oxford Instruments 'Medilog', which can be carried in a flying suit pocket.

2 *Cockpit Voice Recordings*

Not all aircraft are fitted with a cockpit voice recorder, but where this is available, it is useful in three ways. Firstly, it provides a means of identifying the timing of key tasks which can then be related to the continuous heart rate recordings. Secondly, it can be used as a verbal notebook — writing may be impossible during the task under investigation and a verbal record may be the only way to obtain ratings during flight. Finally, the voice recording can provide additional evidence of workload and, although it is not suggested that voice stress analysis should be used in this context, this often prompts the researcher to look more closely at specific elements of the task. The recording can also be used to help the pilot to recall the task during the post flight debrief.

3 *Subjective ratings*

Subjective ratings can be used to give estimates of perceived absolute or relative workload. Absolute subjective ratings are best obtained from ratings made during the flight whereas relative ratings are obtained during the post-flight debrief. The techniques suggested for obtaining subjective ratings for this task are as follows:

3.1 *In-flight ratings* The modified version of the Cooper-Harper Scale which is referred to as the 'Bedford Scale' is the preferred method of in-flight rating. It is a scale which has been validated in a wide number of trials and with many different aircraft types and has been shown to be usable during low-level flight and under extremely high workload conditions (see above report by this author). The single seat task presents two particular problems for the use of this technique. Firstly, the task contains too many individual elements for the pilot to rate and many are of extremely short duration and occur at a time when distraction of rating would be unacceptable. Secondly, the single seat pilot does not have a crew member to prompt him when a rating is due or to refresh his memory if he cannot recall the exact definitions of the scale. In order to overcome these constraints it is suggested that the task is divided into elements which meet the following criteria:

- 3.1.1 They are sufficiently low in number (10 or less) that the rating task will not overload the pilot.
- 3.1.2 They represent elements which are meaningful and which will provide the data to answer the requirements of the study.

SUBDIVISION OF THE TASK INTO ELEMENTS

A flying task can be divided into elements in a number of different ways. The way in which the task is subdivided will have an overriding effect on the data and the conclusions of the study and great attention must be given to this from the earliest stages in the experimental design.

In many studies, flying tasks are divided into those elements which are used for mission planning, ie take-off, checks, climb etc. This may be a valid procedure for some studies however, careful consideration should be given to accepting such a classification.

By using a task classification of this sequential type, it would be possible to monitor the differing workload levels throughout a mission or sortie. However, if one is interested in the difference between psychomotor and monitoring tasks, it may be necessary to use a totally different classification.

In addition, the method chosen for the measurement of workload may constrain the choice of taxonomy. In order to be compatible with the method suggested in this article, for example, the number of subjectively assessed elements would have to be kept to a small enough number to be used in a matrix during the post-flight debrief using the 'Saaty' method. In summary, the subdivision of the task into elements should meet the following criteria:

- 1 The number of elements should be kept to 10 or less to allow the 'Saaty' method to be used during the post-flight debrief.
- 2 The elements should be meaningful to the subject (in this case the pilot) and should provide the data necessary to answer the requirements of the study.
- 3 The temporal spacing of the elements should be such as to permit in-flight assessment.

METHOD

PART 1 - TASK SUBDIVISION

Having defined the data requirements for the study, the Recce/attack task is subdivided into 10 elements, each of which will be assessed in the air and during the post-flight debrief using the 'Saaty' method.

A possible subdivision of the task, which meets the criteria set out above, would be:

ELEMENT	DESCRIPTION
Element 1	Approach to IP. This element includes all of the subtasks for the 3 minute navigation leg to the target: <ol style="list-style-type: none"> a. Checking slip, adjusting speed. b. Weapon switching to final arming. c. Map to ground track check. d. Revision of ETA for IP ± 5 secs. e. Checks of wing mans 6-o'clock. f. Estimate s/w for weapon release. g. Set wind/am depression for attack. h. Ht fix to IP; update pressure alt or auto ht fix at IP (INAS). i. Set or confirm next heading.
Element 2	Acquire IP visually.
Element 3	IP to pull-up. All tasks within this leg should be assessed together.
Element 4	Acquire target visually.
Element 5	Attack manoeuvre, to include the following items: <ol style="list-style-type: none"> a. Top at ht required for dive angle. b. Check speed/power. c. Sight/bomb fall line on tgt. d. Final arming switch (peace time). e. Start camera if not auto. f. Phase change if INAS equipped.
Element 6	Weapon release, including final tgt tracking.
Element 7	Recover from dive.
Element 8	Defensive manoeuvres.
Element 9	Egress, including switches safe and track to next turning point.
Element 10	Locate and identify other aircraft, switch off camera, regain HUD NAV mode.

Some of the elements are summaries of portions of the sorties while others are individual times. Where several items are combined, this portion should be assessed at the end of the leg concerned. The very high workload portion of the task will clearly be during the attack and weapon release phase; the number of elements to be assessed during this phase has been kept to a minimum whilst still retaining the required level of detail and discrimination in the task.

PART 2 - EQUIPMENT PREPARATION

Before the flight, the physiological and voice recording equipment should be checked and tested. If possible the pilot should have the optimum electrode attachment points identified and marked on his chest.

The post-flight 'Saaty' matrix should be typed. The matrix for this task would be 10 x 10 with each element summarised on the side of the matrix.

PART 3 - BRIEFING

It is assumed that the pilot will have been consulted during the preparation of the subtasks which are to be assessed.

However, during the pre-flight briefing, these should be reviewed. The pilot should have been given ample opportunity to rehearse the definitions of the 'Bedford' workload scale and a copy should also be attached to the knee pad.

The experimenter should give a comprehensive briefing on the entire sortie to include:

- 1 Procedures for switching on/off heartrate recorders if not used throughout the sortie.
- 2 Review of all assessed elements to include timing for giving verbal ratings.
- 3 Procedures for giving a verbal commentary which might be useful during the analysis.

PART 4 - DATA COLLECTION AND ANALYSIS

After the completion of the sortie, the pilot should be asked to complete the 'Saaty' workload matrix. This task should be completed as soon as possible after landing but should be preceded by a briefing to refresh the marking method. The experimenter should also be on hand to answer questions regarding the completion of the matrix. Once complete, the matrix should be checked to ensure that it has been correctly filled in.

The pilot should be debriefed and notes taken.

The heartrate recording and the voice tape should be marked to identify the flight. The ground crew should also be supervised to ensure that the flight data recording (where applicable) is removed and sent for analysis. Where possible the analogue heartrate printout and the flight data printout should be on the same scale so that one can be placed alongside the other.

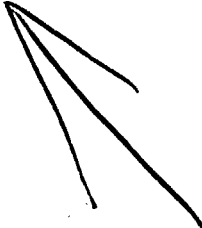
The voice tape should be transcribed and the workload ratings noted. Where there are missed ratings the pilot may be able to provide a rating retrospectively.

The detailed analysis and interpretation of the heartrate recordings is outside the scope of this brief article. The workload ratings from 'Saaty' method can be analysed using the method outlined in the paper above by this author.

SUMMARY

Using a Recce/Attack task as an example, a method for workload assessment has been proposed. The technique relies on the use of subjective ratings scales and physiological measures supported by voice recordings and flight data recordings.

A subdivision of the task has been proposed which permits the use of the 'Bedford' scale during flight and the 'Saaty' method during the debrief. The elements of the sortie have been devised to give the maximum discrimination between the key portions of the sortie while presenting the pilot with a practicable rating task.



CHAPTER 12

IN-FLIGHT ASSESSMENT OF WORKLOAD USING PILOT RATINGS AND HEART RATE

by

Alan H Roscoe
 Britannia Airways
 Luton, England

INTRODUCTION

At present the most used and probably the most reliable methods for assessing pilot workload in flight are based on some form of subjective reporting by experienced test pilots. Unfortunately, subjective opinions are susceptible to bias and pre-conceived ideas and so may occasionally result in false estimates of workload. For more than fifteen years subjective reporting by pilots at RAE Bedford has been augmented by recording their heart rates. At first pilots described workload in a relatively unstructured manner but the need for some form of rating scale was soon apparent. After much trial and error and with the valuable assistance of practising test pilots a ten-point rating scale using the concept of spare capacity was developed (fig 1). The overall design is based on the Handling Qualities Rating Scale of Cooper and Harper (4) already familiar to Bedford test pilots and sometimes used previously, though mistakenly, to rate workload (2).

During the last eight years a number of flight trials at Bedford, including the Harrier 'ski-jump' take-off trial and the Economical Category 3 landing trials, have used pilot ratings and heart rate responses to assess workload (3) (4).

The rationale for using heart rate in assessing pilot workload is based on the concept of neurological arousal. Flying an aeroplane, especially during the more difficult manoeuvres, requires the pilot's brain to collect, filter and process information quickly, to exercise judgement and make decisions, and to initiate rapid and appropriate actions. This neurological activity — which must have been essential for the survival of primitive man — is associated with a state of preparedness sometimes known as arousal. There is evidence that increased arousal up to a moderate level enhances a person's capacity for complex skills; and it has been suggested that the relationship between performance and arousal can be described by an inverted 'U'-shaped curve (5)(6). There is also some experimental evidence that a similar shaped function describes the relationship between performance and task demands. In addition it has also been suggested that levels of arousal are determined by task characteristics or demands, by how an individual perceives the situation, and by how he responds to his environment (7)(8). It is hypothesised that a pilot is more likely to produce an adequate — if not optimum — level of performance by matching his arousal to the perceived demands or difficulty of the flight task. A coarse setting of his arousal may be followed by fine tuning as the task develops. Heart rate tends to reflect neurological arousal via activity in the autonomic nervous system. An appropriate definition of pilot workload, modified slightly from that proposed by Cooper and Harper in the introduction to their Handling Qualities Rating Scale, is: pilot workload is the integrated mental and physical effort required to satisfy the perceived demands of a specified flight task. The interpretation of workload as effort is one that appears to agree with the views of more than 80% of military pilots and civil airline pilots (9), as well as being consistent with the effect on piloting ability of a number of individual variables.

Description of the Technique

Workload ratings — It is almost essential when using a workload rating scale to specify the flight task in reasonably precise terms. The workload being assessed should be that involved in the execution of the primary task. The pilot will almost certainly be performing additional tasks, but the effort expended on them must be included as part of his spare capacity.

Ratings, which should be given in flight wherever possible, may be for a complete flight task, for example, an instrument approach and landing, or for a sub-task, such as becoming established on the glide slope. On the other hand an experimental protocol may require regular ratings at specified time intervals which might vary according to the stage of flight; perhaps being more frequent during expected high workload phases of flight. Regular ratings of this kind tend to be less reliable unless related to a particular flight task.

The rating scale is not linear and probably lacks sensitivity at the lower end; half ratings are allowed within each decision branch and tend to be used frequently. Originally it was decided not to permit the use of half ratings between the decision branches but the occasional difficulty of deciding between the last two branches, in effect between ratings 3 and 4, was resolved by accepting a rating of 3½.

It is important that pilots are fully briefed on the scale to be used. In its final form this particular scale has been generally welcomed by pilots who find it relatively simple to use in practice, especially so if the task to be rated is short and well defined. Somewhat surprisingly, a few pilots unfamiliar with rating techniques have recently used the scale with good effect in assessing workload on Boeing 737 and 767 aircraft. These favourable observations are probably due to the use of a definition of workload accepted by pilots and to basing the scale on the idea of spare capacity.

Recording Heart Rate Heart rate recording is non-intrusive and it is compatible with flight safety; pilots seem readily to accept being 'wired up'; and the discrete nature of the basic data encourages various forms of analysis. The technique used to record heart rates from pilots during flight is based on the electrocardiogram (ECG). Amplified ECG signals, detected by means of two disposable electrodes applied to the pilot's chest, are recorded in analogue form on magnetic tape along with speech (which might include workload ratings) and, where possible, other aircraft parameters. In the first instance the basic signal — the 'R' wave of the ECG — is plotted out along with heart rate in instantaneous or 'beat-to-beat' form (derived from the 'R' waves by cardiostimulator). Subsequently mean rates for a particular task, sub-task, or time interval may be calculated according to the requirements for workload ratings. Plots of mean rates for 30 sec epochs are often useful in demonstrating

AD-P005-639

PILOT WORKLOAD RATING SCALE
(for a specified piloting task)

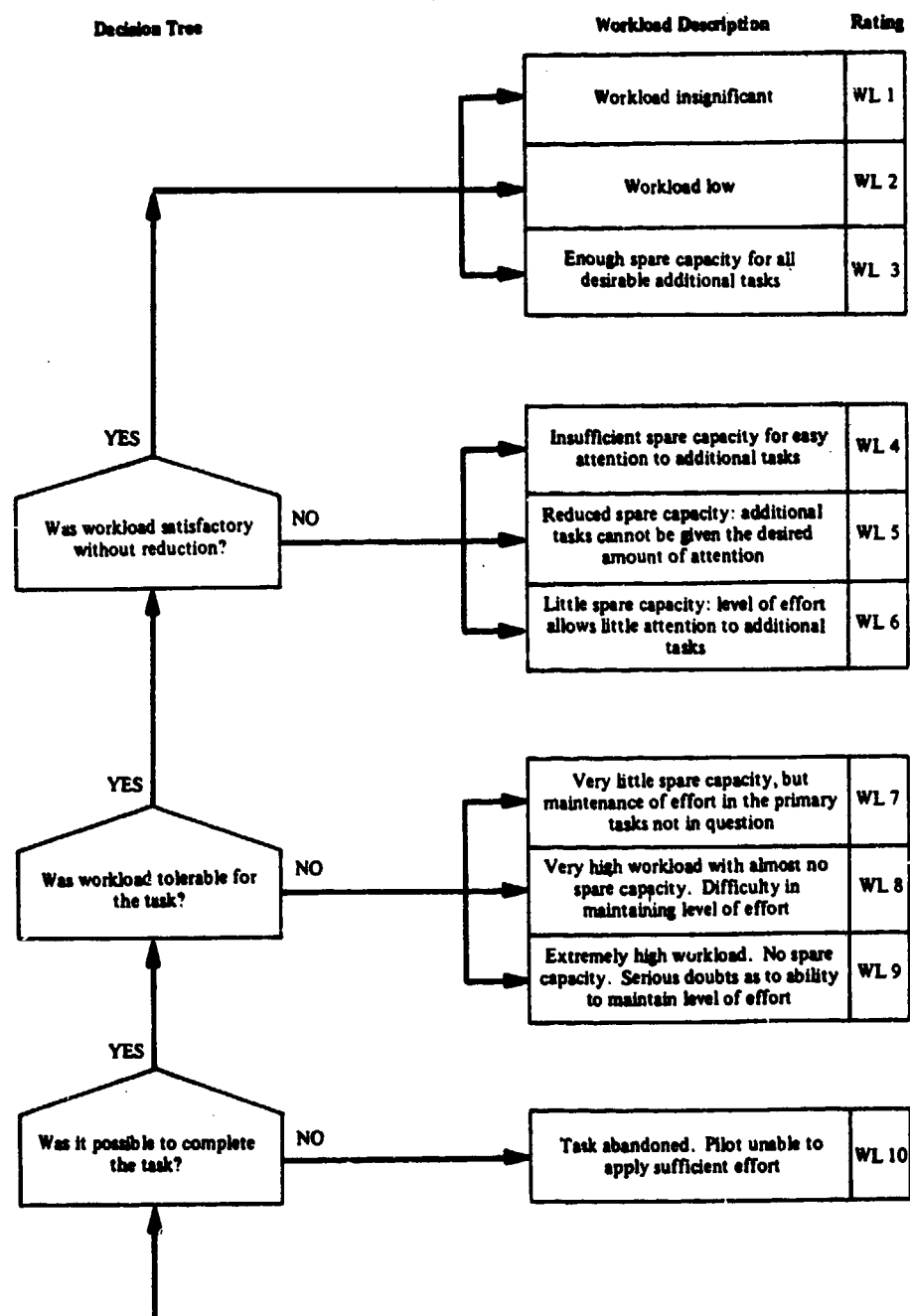


Fig.1 Pilot workload rating scale
The decision-making process is started at the bottom left corner of the 'decision tree'

*The workload being assessed is that involved in the execution of the primary task. The pilot will almost certainly be performing additional tasks, but the effort expended on them must be included as part of his spare capacity.

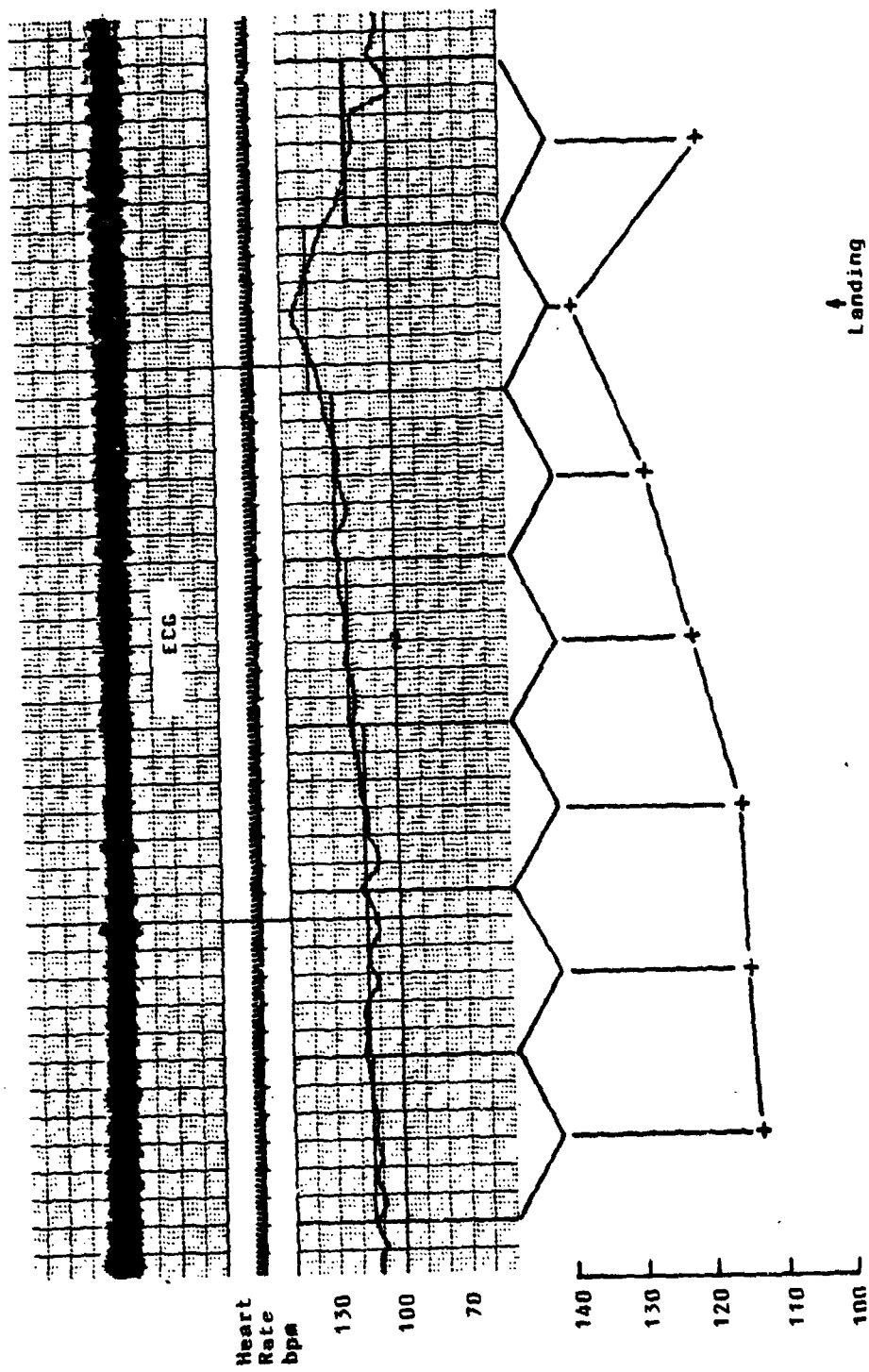


Fig. 2 Schematic representation of a 30 sec plot derived from beat-to-beat heart rate; approach and landing B767

significant heart rate changes by smoothing (fig 2 is an example). On the other hand, beat-to-beat plots have the advantage of showing rapid and sometimes short term changes of interest.

In the absence of any significant change in overall heart rate the degree of sinus arrhythmia (physiological heart rate variability) may be of value in assessing changes in mental workload. Changes in sinus arrhythmia are usually evident on visual inspection of beat-to-beat plots; a number of techniques are available for scoring sinus arrhythmia although none seem to be reliable and so results must be interpreted with caution.

Performance — As workload and performance are to a large extent interdependent it is important when assessing the former to monitor the latter. In some flights it is a relatively simple matter to record actual performance in the air by means of aircraft recorders or on the ground by kinetheodolites. Where this is not practicable realistic performance limits should be defined and monitored by a flight observer, by video recording, or by the pilot himself.

Example of using the Technique

Assessing pilot workload during a manually flown instrument approach and landing using a flight director system in a twin-jet transport flown by a crew of two pilots. (See Appendix I for details). The defined task lasts five minutes.

Heart rates are recorded from both pilots continuously throughout. Workload ratings are requested from both pilots and for each pilot from an experienced flight observer seated on the flight deck as follows:

1. At 3,000ft — starting the final descent onto the glide slope.
2. At 1,000ft QFE — for glide slope acquisition.
3. At 100ft QFE — for final approach.
4. On deceleration to 60K — for flare and touchdown.

Upward events are rated on an *ad hoc* basis. Performance is monitored by the flight observer. Mean heart rates for the appropriate periods before the ratings are calculated and bracketed with the rating scores. The beat-to-beat heart rate plot is examined for evidence of inappropriate or sudden changes and also for suppression of the sinus arrhythmia. (Inspection of heart rate plots by the pilots will often act as an *aide memoire*). Ambiguities and inconsistencies are of particular interest and are studied in more detail.

These data provide some idea of workload levels but become more valuable when compared with data from the same pilots recorded on other occasions when using different techniques or systems, or when flown in different weather conditions. For example, this flight director approach may be compared with an approach using a different type of flight director, with a raw ILS approach, or with an autoland.

Pitfalls and Limitations

The technique described above does not result in the more precise measurements associated with experiments carried out in the controlled conditions of laboratories. Furthermore, there are a number of important limitations and pitfalls to be aware of when assessing levels of workload in real flight.

1. Ratings depend largely on the personal experience of the pilot and do not result in absolute values of workload, comparisons between pilots are, therefore, not valid; minor inconsistencies between different pilots flying the same aeroplane should be expected.
2. In-flight ratings may not be possible when assessing workload in single-seat aircraft.
3. As the rating scale is non-linear statistical treatment of rating numbers must be treated with caution.
4. The idiosyncratic nature of the heart rate response precludes comparison of results derived from different pilots — each pilot must be used as his own control — unless large numbers of pilots are involved.
5. Heart rate responses recorded during flight tasks involving increased physical effort or physical stressors such as high 'g' manoeuvres must be interpreted with care.
6. Ambiguities and inconsistencies between a pilot's ratings and his heart rate responses are sometimes due to a pilot rating a particular aspect of part of a task or epoch rather than the entire task or period of time.
7. The technique is most valuable when the handling pilot is manually flying the aeroplane during a relatively demanding task or when he is anticipating taking manual control at short notice. Both ratings and heart rate responses for non-flying pilots in a purely monitoring role are less valuable, although changes in beat-to-beat heart rate variability can be most useful in detecting changes in mental load.
8. Finally, experience so far suggests that results from one pilot in five show poor agreement between subjective ratings and heart rate responses. The reason for this disagreement is not known for certain but may be due to the failure of heart rate to reflect accurately levels of central arousal in these individuals.

REFERENCES

- 1 COOPER G E
HARPER R P The use of pilot rating in the evaluation of aircraft handling qualities. NASA Technical Note 5153, Washington DC, 1969
- 2 ELLIS G A Subjective assessment pilot opinion measurements in: Roscoe A H (Ed.) Assessing pilot workload, AGARDograph AG-233, AGARD Paris 1978
- 3 ROSCOE A H Assessing pilot workload in flight in: Conference Proceedings No 373, Flight test techniques. AGARD Paris 1984
- 4 ROSCOE A H Pilot workload an Economic Category 3 landings in: Conference pre-prints. Aerospace Medical Association Annual Scientific Meeting. 1980
- 5 DUFFY E Activation and behaviour. J Wiley and Sons New York 1962
- 6 HOCKEY R Stress and the cognitive components of skilled performance in: Hamilton V and Warburton D M (Eds) Human stress and cognition. John Wiley and Sons Chichester 1979
- 7 WELFORD A T Stress and Performance. Ergonomics 16 567-580 1973
- 8 KAHNEMAN D Attention and Effort. Prentice - Hall Inc Englewood Cliffe 1973
- 9 ELLIS G A
ROSCOE A H The airline pilot's view of flight deck workload: preliminary study using a questionnaire. Royal Aircraft Establishment Technical Memorandum. FS(B) 465 1982



CHAPTER 13

THE ASSESSMENT OF WORKLOAD IN HELICOPTERS

by

Helen C Muir and Robert Elwell
Applied Psychology Unit
College of Aeronautics
Cranfield Institute of Technology
Bedford MK43 0AL, England

The value of inflight assessment of pilot workload has been recognised by aviation researchers and designers for over a decade (1) (2). Initially the subjective reporting of workload by experienced test pilots was based upon an application of the Handling Qualities Rating Scale of Cooper and Harper (3). This subjective reporting led to the development of rating scales for the assessment of workload (4). These subjective techniques were later augmented by the recording of physiological variables which could be interpreted as indices of workload (5) (6).

In the last decade, rather than restrict the assessment of workload in aviation to data obtained from test pilots, studies have been reported in which small samples of professional pilots have been used (5) (6). A more recent development has been the employment of workload measures for exploring differences between pilots and to look for correlations between these measures and performance, and success in training (7). Workload estimation has additionally been used to assist in the ergonomic design of systems including crew station geometry, and control and display location (8) (9).

In these, and other cases, the requirement to measure workload has had a practical and 'applied' character. It results from a need to specify and predict the future performance of the operator within a system; to determine what effect will result from changes to an existing system, or evaluate the consequences of entirely new procedures or technology. To this extent workload is fundamental to a wide variety of disciplines.

Although there is broad agreement on the importance of workload, partly as a consequence of the wide range of areas to which the workload concept may be applied, there is no universally agreed definition. In any investigation in which an assessment of workload is to be made, a definition will obviously be required as a basis for both briefing subjects and interpretation of the results. A definition which is frequently used in both aviation and other areas is "the combination of physical and mental effort required to complete the task".

Workload concepts may in fact be refined into 'physical' and 'mental' subsets, represented at extremes by the power output of manual workers to studies of 'decision making' (10). The pilot's task is a combination of the two, with advances in technology emphasising the mental element, ie, monitoring, anticipating, decision making, the need for the pilot to wrestle with the flight controls is largely dated. However, for the military pilot these same technological advances are tending to degrade the physical conditions under which performance is required, eg, increased g, thermal changes, longer duration sorties, more restrictive (albeit more efficient) protective assemblies. Similarly, the air transport pilot encounters more sectors in a duty period, or more rapid change of time zone.

Besides the approaches of different disciplines to the investigation of workload in aviation, there are two other conditions, which at a fundamental level, are extremely difficult to isolate from workload; these are stress and fatigue (10). The concepts if not defined in terms of one another, are implicitly inter-related. Thus if workload is defined in terms of effort (as above), such expenditure cannot be continued indefinitely, hence fatigue. Increased workload will therefore imply the faster onset of fatigue. In turn the mental and physical concomitants of exhaustion may be characterised as stress. Stress results from an excessive demand on the individual.

Workload studies may be employed to determine the current and potential operating capacity of a system. It may be that the material assets are fixed, by that re-scheduling, or re-rostering of crews, or re-defining their duties can allow greater efficiency. For the military, an aim may be to achieve greater combat efficiency, whilst in civil aviation it may be to take on extra routes or services. Other objectives, which are not exclusive may include, increased reliability, efficiency or safety.

Thus workload assessment is frequently a component in a programme with externally defined objectives which tends to follow a particular pattern.

The stages which might be required for a programme of research, and the reasons for their inclusion can best be described by reference to a specific study. One study of this nature, currently being conducted by the authors, is to determine the appropriate allocation of tasks between two pilots manning an Army helicopter.*

The research is undertaken in 5 discrete steps. These are described below and summarised in Figure 1.

STAGE 1: DEFINITION OF PROBLEM AND RESEARCH OBJECTIVES

When any programme of research is required, prior to the actual commencement of work, there is the rather obscure phase of the organisation requiring the research (sponsor) coming together with the researchers (who may be internal or external to the organisation). Initially it may be difficult for the organisation to recognise the true nature of problems which may

* The Army Personnel Research Establishment (UK) have commissioned the College of Aeronautics at Cranfield Institute of Technology to carry out a programme of research in order to determine task allocation between helicopter pilots and to develop Standard Operating Procedures.

AD-P005 640

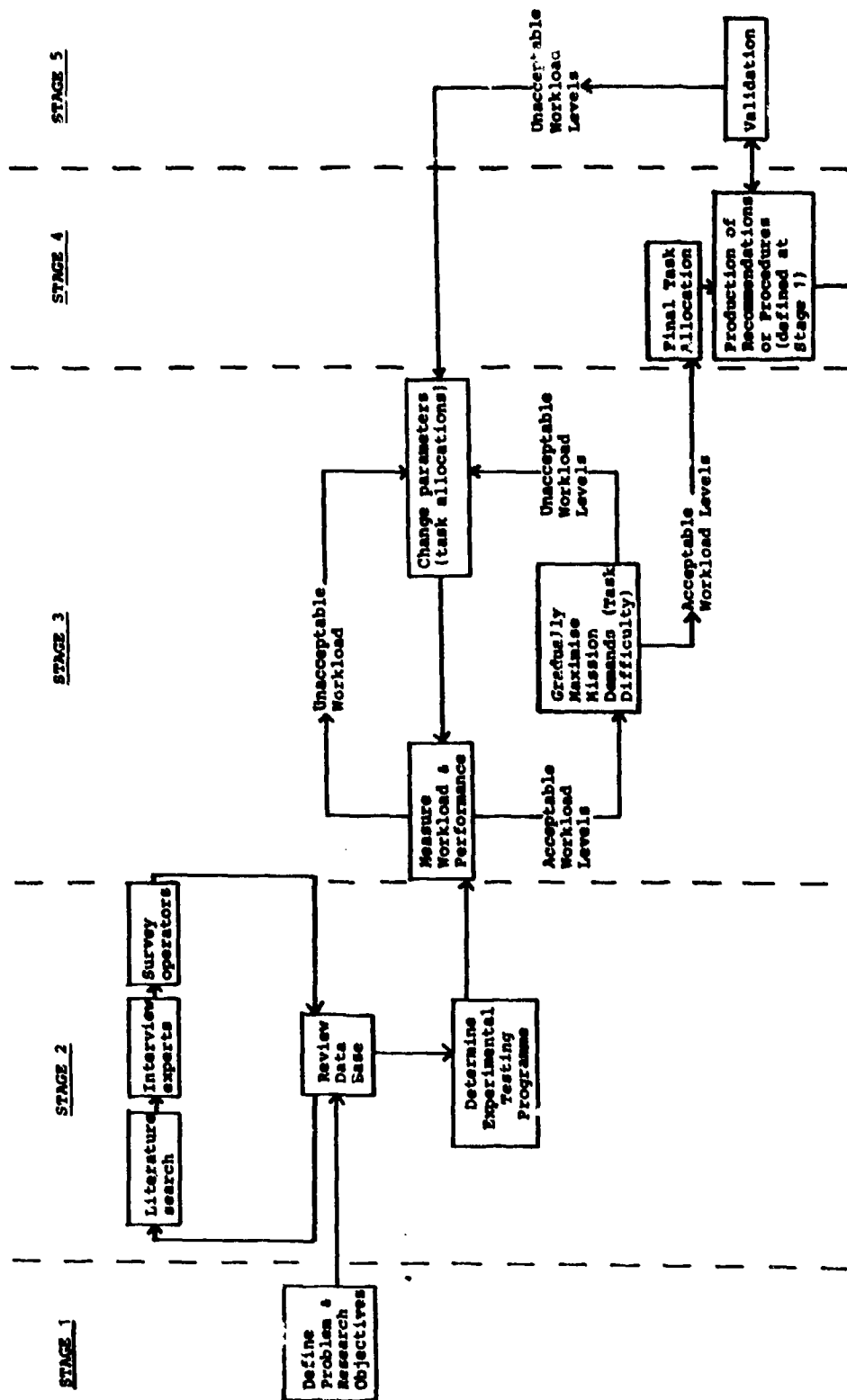


Fig.1 Summary of research stages

have arisen over time or be generated by operational change. The organisation may not be aware of the contribution human factors research and the benefits of rigorously applied techniques can make to resolving such changes; and confusion is likely to exist about what constitutes an optimum resolution.

The ideal case is when an organisation either recognises an existing problem, or foresees one in the future. It is important to realise that the problem may range in scale from a high accident rate through poor crew communication, air traffic procedures, to rostering. The problem may not only involve the flight crew. A problem becomes amenable to human factors research, and particularly to workload assessment techniques, when there is an involvement of people in the process and it is believed that it is the performance of these people which is the limiting factor.

The sponsors, having identified a problem, accepted the need for behavioural analysis, and engaged researchers, must define their objectives. These are the objectives by which a solution derived from workload assessment techniques may be judged. Whilst this is of critical importance to the validation of the study (discussed below), it has the additional advantage of forcing the sponsors to consider fully the implications of their identification of a problem.

In the helicopter study currently being undertaken by Cranfield, as part of Stage 1 it was agreed with the Army Air Corps that the primary objective of the research would be to determine the appropriate allocation of tasks between two pilots manning an Army helicopter. This would be derived by separately assigning flight and combat tasks to crew members. The emphasis would be upon the operational employment of the helicopter. This would be achieved by analysing crew workload during flight and subsequently determining which crew member would best perform which tasks. Analysis of current and projected mission profiles would be undertaken to determine how these affect task allocation between the crew. Finally, Standard Operating Procedures would be drafted.

The complexity of the aircrew task together with the need to reproduce with maximum fidelity and conditions prevailing whilst actually flying at ultra low level meant that objective evaluation of aircrew workload had to be taken in the air with representative missions. The use of simulators and non-aircrew subjects of "equivalent" tasks was not considered to be sufficiently representative.

STAGE 2: REVIEW OF THE DATA BASE AND DETERMINATION OF THE EXPERIMENTAL TESTING PROGRAMME

Included in the concept of the Data Base are the abilities, skills and techniques of the researchers themselves, the information that can be derived from the appropriate academic, and organisational literature, and knowledge that is held within the organisation itself. For instance, management may be aware of a problem, but unclear about the details of related processes, these being the province of experts.

This in turn can contribute to the difficulty in identifying a problem: for instance when senior pilots are promoted into management positions this can occur, either because they are cushioned by their status from everyday operations or the system has evolved subsequently.

On the flight deck, the experts are the instructors and training captains. Detailed individual interviews with them will increase the researchers understanding of how the operations proceed, and normally provide clear insight into the scope of the problem under investigation. The other major source of information will be the ordinary flight crews. For this large group a survey of opinion, by questionnaire, is frequently the most appropriate method of data collection.

It is clear that the ways in which the Data Base can be refined are as varied as the operations that are under investigation. Also that the depth of analysis required is variable, whilst the sources of information that could be consulted, freed from constraints of time or cost, are virtually unlimited. The duration and extent of this phase is therefore dependent upon the research team's prior knowledge. Ideally, the research team should include a psychologist and a pilot.

In the helicopter study, this stage involved a literature search and a series of informal and semi-structured interviews with experts in aircrew training, tactics, standards and safety from the Unit which had the requirement for the investigation and who were responsible for the operation of the missions.

The technique of semi-structured interviews involves the use of sequentially structured general questions which lead to choice or branching questions. Having registered a preference in response to a particular question the interviewee is then asked to describe the reasons for their choice. The interview may be recorded on tape in order that the responses from all of the interviewees may be pooled and used to provide information for subsequent stages of the research.

The importance of these interviews should not be underestimated since without their inclusion assumptions may be made regarding organisation and deployment in the operational unit based exclusively on the beliefs of the commissioners of the study and the researchers. Data from this stage provide information regarding current and proposed mission profiles and an exhaustive list of potential crew tasks (and potential allocations) together with priorities and as assessment of criticality to mission success.

This stage also involves the construction of a questionnaire based upon the results of the interviews asking for subjective ratings of workload for the tasks identified on representative missions. This is applied to current aircrew members and to the experts who initially provided the information. This should be supplemented within a small percentage of the former group by short informal interviews. Stage 2 strengthens the representativeness and validity of the data collected.

As part of Stage 2 a representative sample of aircrew who will be required to participate in Stage 3 is determined. Studies are reported in the literature, especially regarding workload and cockpit assessment in which the sample was limited to a number of pilots who were unlikely to represent the full range of the user population.

Helicopter pilots may be called upon to undertake an almost infinite variety of different sorties, each of which will impose a particular load upon the crew. To utilise workload measurement techniques effectively requires that this variety is reduced into a set of standardised (and hence reproducible) profiles. Within these profiles the facility must be available to change the loading on different crew members, and a promising and quantifiable technique is to vary the communications load according to the level of difficulty that is desired.

Derivation of suitable profiles in the helicopter flight regime follows from the analysis of the aircraft tasks (via interviews and workload questionnaires) and a study of the ways in which such tasks are carried out. This information may be obtained from the Operations Manual of a civil company, or the Tactical Doctrine promulgated by a military operator. Shown in Figure 2 is a 'Simple' mission which might be undertaken by a reconnaissance helicopter of the British Army Air Corps. The task briefed might be to look for the enemy, from a defined geographical area, and to report any sightings. The transit to the area can be made more or less demanding of map-reading and low flying skills by imposing "realistic" constraints, such as imaginary artillery positions, drone launches. Radio communications can be required to increase the cockpit activity.

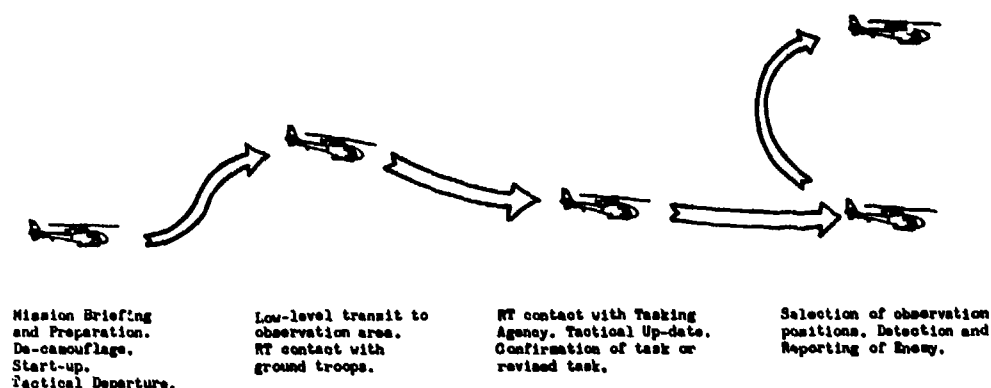


Fig.2 Hypothetical 'simple' mission for reconnaissance helicopter

The selection and occupation of observation positions is a demanding activity in itself, whilst detecting, and subsequently locating the enemy on a map can be varied by using actual vehicles to provide a real target. A further performance measure may involve the 'enemy' using a video system to record occasions when the reconnaissance helicopter is visible, and hence vulnerable, ie occasions of poor performance.

This 'Simple' mission can be extended (as demonstrated in Figure 3) merely by requiring the aircrew to complete the sortie. Thus the helicopter can be relieved by another (probably a notional one), and the crew must be briefed upon the current situation, whilst on the route back the tasking agency must be updated, and further navigation hazards can be introduced.

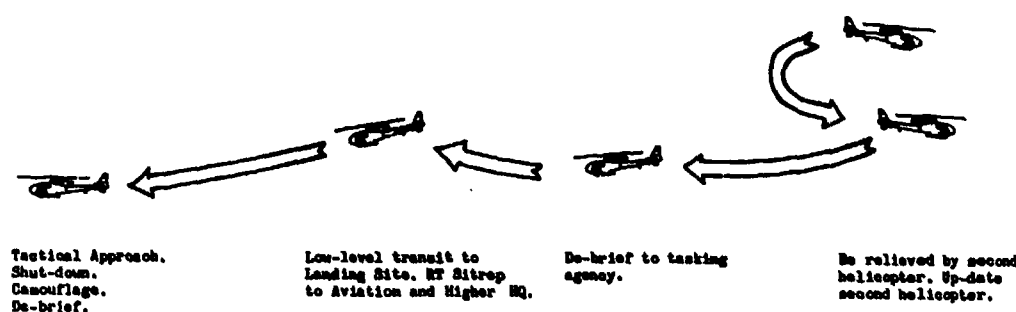


Fig.3 Extension of the 'simple' mission

A major increase in activity can be introduced by requiring the aircrew to make a more comprehensive identification of the enemy, and then assess its suitability for attack by various different weapon systems, eg artillery or Fighter Ground Attack aircraft. In the example in Figure 4 the appropriate system is Anti-Tank Helicopters. By the integration of other aircraft on normal training exercises, into the experimental sortie, the aircrew task can be made as realistic as possible.

Select observation positions.
Manoeuvre to view enemy.
Detect enemy.
Identify enemy.
Assess direction and rate of movement.
Report to tasking authority

Maintain observation by changing observation positions.
Recommend suitability for Helarm.

Maintain observation by changing observation positions.
Accept tasking as Helarm Director.
Select Rendezvous with Anti-Tank Helicopters, and choose probable Engagement positions.

Possibly assume command over second recon helicopter.
Brief and task.

Brief Anti-Tank helicopters at rendezvous. Obtain attack clearance from tasking authority.

One recon helicopter leads Anti-Tank Helicopters to fire positions (and guards flank) other engages enemy with artillery.

Having fired, Anti-Tank helicopters depart. Recon helicopter reports damage. Selects new position.

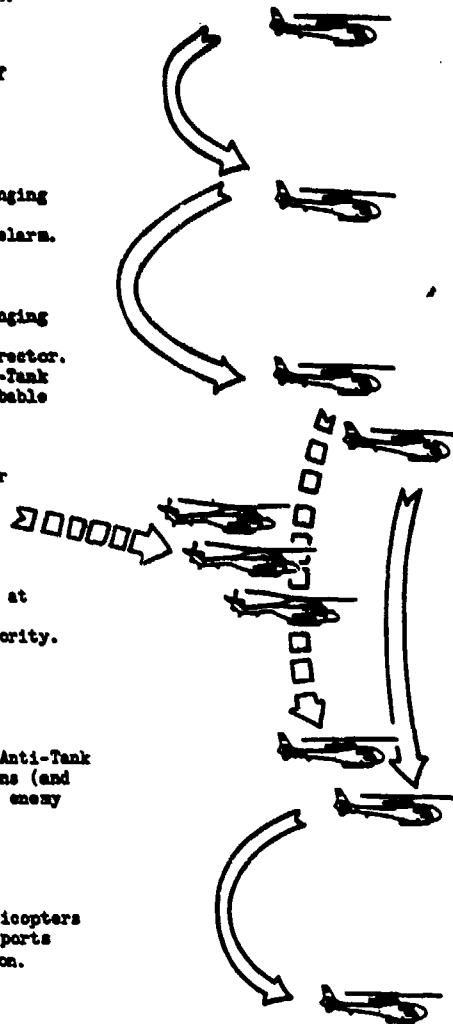


Fig.4 Possible increase in mission difficulty

STAGE 3: MEASUREMENT OF WORKLOAD AND PERFORMANCE

While the variety of techniques devised to study workload is extensive (11) those that may be applied on the flight deck or in the cockpit are few in number, and the methodological and technological restrictions cause severe constraints. Most significant are the safety implications of equipment, compatibility with aircraft systems, and potentially hazardous interference with crew activities (12). Already perhaps encountered in the interview and questionnaire stage may be a misapprehension about the purpose and outcome of the research. From the individual's point of view these fears may be well grounded if, for example, the workload study was to determine whether two crew could do the work previously done by three (13).

This observational stage would involve the inflight assessment and workload measurement of the aircrew task. A member of the research team should fly with typical crews on representative missions determined at the end of Stage 2.

The workload data collected from the aircrew on these sorties falls into two categories:

- Subjective ratings of pilot workload. It is important that a definition of workload is agreed by the research team and sponsors prior to the commencement of the study and that this definition is given to the crew as part of their briefing before the mission sorties commence. The workload rating used in the helicopter study is based on the Bedford Scale (14).
- Psychophysiological indices of workload. Although the most widely used method for assessing workload is that of subjective rating, since it cannot be directly observed, the effects of workload may also be inferred from differences in measurable physiological functioning. There is a wealth of evidence that subjective ratings of workload may be usefully augmented by certain physiological indices (5) (15).

The selection of suitable measures will be influenced not only by the theoretical requirement for particular data but also by the feasibility of collection. Agreement between the various measures is frequently tenuous. This results from both the difficulty of obtaining reliable readings, and relating these to real-world events. To a considerable extent the restrictions that must be placed upon data collection in aircraft constrain the researchers' choice of methodology. Comprehensive reviews of the issues may be found in O'Donnell (16) and Wierwille (15).

In the helicopter investigation it was decided to limit the recording of physiological data to that of cardiac activity. The reasons for this choice were that the data can be reliably collected from both members of the aircrew with minimal interference and that such measures have been reported by most researchers as reflecting, to a considerable extent, cognitive as well as physical activity.

From the cardiac data two independent indices may be derived.

- 1 Mean Heart Rate — the instantaneous heart rate derived from measured time intervals between successive ventricular contractions (R-waves of the cardiogram) expressed as beats per minute.
- 2 Heart Rate Variability — this takes into account the normal physiological trend of mean heart rate and minimises its effect by considering observation order through difference scores.

Although mean heart rate and heart rate variability have been shown to correlate highly with subjective indices of workload, there is evidence to suggest that mean heart rate can be more valuable in situations of high workload whereas in situations of relatively low workload then heart rate variability may be the most sensitive measure.

In any task analysis or assessment of workload it is essential that both subjective and physiological indices of workload are related to objective measures of performance.

The experimenter must have criteria to assess aircrew performance, because it is when performance falls below the specification that one may say that workload is excessive. In the relatively straightforward assessment of performance in, for example, fixed wing public transport aircraft, accuracy in maintaining flight parameters might suffice. In contrast, if the speed height and direction of the combat helicopter are not changing the experimenter must suspect overload.

Video pictures of cockpit activity provide the most accurate and objective recording of events, particularly when compared to either observer or the subject's own reports. In the helicopter study video cameras will be mounted in the cockpit of the helicopter and will be used to provide video recording of the cockpit activity and crew interaction. An additional advantage is that the record is permanent and this will allow the re-analysis of early sorties as the data from subsequent ones is obtained. An adaptation of the methodology developed by Lovesey will be used for the analysis of the video recording.

The member of the research team who is trained both as a pilot and a psychologist, will fly as an observer and record gross crew activities as the sortie progresses. The intercom and radios will be taped to provide a record of crew interaction.

Analysis of the performance data from the observer reports, as well as the recordings from the video, intercom and crew radios will allow an accurate record of crew activities to be compared against objectively specified mission parameters. These parameters will have been developed from the data obtained from Stage 2.

Cardiac recordings time locked to the activity record will permit the independent assessment of pilot workload. This will be correlated with the aircrews' own assessment of the effort involved. It will then be possible to estimate and compare the workload of the crew at succeeding intervals of the flight, and to relate physiological and behavioural measures to mission elements. The recordings may also highlight occasions of under or over loading either crewmember.

Consideration was given to instrumenting the aircraft flight controls or recording flight path data, however a number of factors led to the rejection of this suggestion when using workload techniques to determine crew loading. Not the least of these was the sheer quantity of data which would be collected. Should the data be collected, it may be of limited use, as skilled performance, especially motor performance, does not decline steadily under increasing workload, rather it continues relatively unchanged until catastrophic failure. In addition only the performance of the handling pilot would be recorded.

If the results from the task analysis indicate either decrements in performance or unacceptably high levels of workload, it will be necessary to change a parameter within the operational flight setting, and repeat the observational process. For reasons of safety, the experimental technique will be to increase task difficulty with successive sorties. This will be achieved by increasing the frequency of task related activity.

STAGE 4: PRODUCTION OF RECOMMENDATIONS AND PROCEDURES

When the test conditions are satisfied the research team should specify revised or new procedures or recommendations. These should resolve the problem and meet the objectives initially identified by the project sponsors at Stage 1. In the helicopter study, this will be the appropriate allocation of tasks between two pilots manning the army helicopters and the development of Standard Operating Procedures.

STAGE 5: VALIDATION

The importance of testing the procedures or recommendations derived from Stage 4 on an independent group of subjects for the purposes of validation cannot be overemphasised. In the helicopter study this will be done by replicating certain of the mission sorties using crews from squadrons which were not involved in the original experimental programme. A representative sample of crews would be required to carry out a number of missions which had been included in the original test programme in Stage 3 and to follow the newly developed Standard Operating Procedures. Performance and workload data would be collected throughout the sorties. Any findings of unacceptable levels of workload or performance would indicate a requirement

to revert to a Stage 3 and carry out a second phase of observations. Finally, the findings from the research programme should be checked against the objectives determined at Stage 1.

SUMMARY

In aviation an assessment of workload is frequently used as one component in a programme of research. The objectives of the research may vary from an assessment of the activities of the crew to an evaluation of either cockpit modifications or operational changes. Thus workload assessment will form one of a series of stages in the research. A model is presented in which the stages of the investigation which will proceed and follow the workload assessment are described. An application of this approach to the assessment of workload in helicopters is used to illustrate the practical implications of the model.

REFERENCES

- 1 HOWITT J S Flight-deck workload studies in civil air transport aircraft. In: Conference Proceedings No 56 Measurement of Aircrew Performance. AGARD, Paris, 1969
- 2 GARTNER W B
MURPHY M R A critical survey of concepts and assessment techniques. NASA Report No NASA TN D 8365 1976
- 3 COOPER G E
HARPER R P The use of pilot ratings in the evaluation of aircraft handling qualities. (NASA Ames Technical Report NASA TN D 5152). Moffett Field, CA: NASA Ames Research Center. 1969
- 4 REID G B
SHINGLEDECKER C A
EGGEMEIER F T Application of conjoint measurement to workload scales development. Proceedings of the Human Factors Society Annual meeting. 1981
- 5 ROSCOE A H Stress and workload in pilots. Aviation Space Environ Med 49, 630-636 1978
- 6 KOCH C
MONESI F Evaluation of aircrew fatigue during operational helicopter flight mission. In: Conference Proceedings No 255 Operational Helicopter Aviation Medicine. AGARD Alabama 1978
- 7 DAMOS D Residual attention as a predictor of pilot performance. Human Factors 20 435-440. 1978
- 8 LOVESEY E J Human factors evaluations of today's helicopters as an aid to future systems designs. In: Conference Proceedings No 225 Operational Helicopter Aviation Medicine AGARD Alabama 1978
- 9 SMIT J
WEWERINKE An analysis of helicopter pilot control behaviour and workload during instrument flying tasks. In: Conference Proceedings No 225 Operational Helicopter Aviation medicine AGARD Alabama 1978
- 10 MORAY N Subjective mental workload. Human Factors 25 25-40 1982
- 11 MORAY N Mental workload: Its Theory and Measurement London Plenum 1979
- 12 CASALI J G
WIERWILLE W W On the measurement of pilot perceptual workload: a comparison of assessment techniques addressing sensitivity and intrusion issues. Ergonomics 1984 27 (10) 1033-1050
- 13 LERNER, E J The automated cockpit. IEEE Spectrum 20 57-62 1983.
- 14 ROSCOE A H Assessing pilot workload in flight In: Conference Proceedings No 373 Flight test Techniques AGARD Paris 1984
- 15 WIERWILLE W W "Physiological measures of aircrew mental workload" Human Factors 21 (5) 575-593 1979
- 16 O'DONNELL R D Contributions of psychophysiological techniques to aircraft design and other operational problems. Conference proceedings No 338 AGARD 1983.



CHAPTER 14

ASSESSING WORKLOAD FOR MINIMUM CREW CERTIFICATION

by

J J Speyer and A Fort
Flight Division, Airbus Industrie
B.P.33
31707 Blagnac Cedex, France

Dr J P Fouillot
Laboratoire de Physiologie
Faculté de Médecine, Cochin
24 rue du Faubourg St Jacques
75104 Paris, France

and

R D Blomberg
Dunlap & Associates East
17 Washington Street
Norwalk, Connecticut 06854, USA

PART I

STATIC WORKLOAD ANALYSIS AND PERFORMANCE ANALYSIS

INTRODUCTION

The verification of both functional effectiveness and human welfare has evidently always been a major objective in flight test but a formal and rigorous investigation of the man-machine interface itself was gradually prompted by the crew complement question.

When the FAA eliminated the 80,000 lbs rule in 1964 it stated implicitly that the weight of an aircraft or its number of engines had no true bearing on whether a third crew member should be included. Instead the FAA adopted rules to base crew complement certification on workload.

The new rule, FAA's FAR 25.1523 on Minimum Flight Crew and its Appendix D (Figure 1), provided a set of design-related, operational and human factors parameters.

In 1979, AIRBUS INDUSTRIE launched a preliminary version of the Forward Facing Crew Cockpit for installation on the A300 FF. This version incorporated all the new technology for the A310 cockpit except the Cathode Ray Tubes for flight data and system monitoring. In view of the two-man crew certification of this version we launched a major workload research programme in order to develop refined, rational and scientific methodologies for workload determination in flight test. This human factors activity was however not just started for certification purposes but also because crew workload is a fundamental design parameter influencing the cockpit design itself and the operating procedures. In this sense our work started a process which potentially could become iterative in the future so that the man-machine interface would eventually be designed, investigated and improved well before an aircraft's first flight.

The critical importance of man-machine interaction has long been recognized in the field of aircraft handling qualities. What is relatively new, however, is the recognition that man-machine interaction is part of a complex information transfer process between the pilots, the aircraft and ground facilities (1). Classical are the systematic methods for assessing aircraft handling qualities such as the Cooper-Harper scale and this even inspired our approach to workload assessment.

Also classical topics in flight test are the determination of static and dynamic stability, the former indicating the tendency of an aircraft to return to its equilibrium position, the latter indicating the way an aircraft returns to its equilibrium position. Analogous to the complementarity of these evaluations, we developed the Static Taskload and the Dynamic Workload Methods which were first used for the two-person crew certification of the A300 FF in early 1982. As shown in Figure 1, these are complementary but overlap in certain areas. Both methods address particular workload functions and factors listed in FAR.25.1523 Appendix D simplifying the verification of results against specific requirements.

WORKLOAD AS A STUDY ITEM

1. Operational Classifications

The vast literature on workload reveals an unusual diversity in the way workload has been defined and used. Clearly there seems to be no generally accepted definition and there is no universal metric, yet no direct method for measurement.

A survey of the literature (2) (3) indicated however that the many operational definitions adopted by research can be gathered into three functionally related attributes, namely *input load*, *operator effort* and *output result*.

REF	FAR 25.1523 Appendix D	APPLICABLE METHOD
a)	BASIC WORKLOAD FUNCTIONS	
1	Right path control	D
2	Collision Avoidance	D
3	Navigation	D
4	Communications	D
5	Operation and Monitoring of engines and systems	S, D
6	Command Decisions	D
b)	WORKLOAD FACTORS	
1	Ease of Operation of flight, power and equipment controls	S
2	Accessibility and visibility of instruments and failure warnings	S
3	Number, urgency and complexity of operating procedures	S
4	Mental and physical effort involved in normal, abnormal and emergency situations	S, D
5	Extent of systems monitoring required en-route	D
6	Actions necessitating unavailability of a crew member at his assigned duty station	S, D
7	Degree of automation provided in the aircraft systems	S, D
8	The communications and navigation workload	D
9	The possibility of increased workload associated with any emergency that may lead to other emergencies	D
10	Incapacitation of a flight crew member	D

Figure 1
FAR 25.1523, Appendix D

Input load or taskload considers workload as a set of observable or identifiable task demands. Task demands are defined for a given scenario in terms of task elements, their nominal time duration, their inherent difficulty and their schedule and requirements i.e. a normative, detailed description of what is overtly required or demanded of the operator or pilot in the performance of a task but it does not measure the resulting physical or mental response of the operator.

The *operator effort or workload viewpoint* is on how hard an operator or pilot must work to satisfy a specified set of task demands i.e. workload is addressed from the standpoint of measurement of covert or internally generated responses to these task demands. Due to the complexity and covert nature of mental functions such as information acquisition, processing and decision making, there is a lack of knowledge about the nature of mental workload. But when we speak of workload we certainly mean something to do with a sense of mental effort, how hard one feels one is working. In general it can be said that mental workload is some undefined combination of mental effort and emotional stress in response to task demand (4) (5). Moreover, a wide range of physiological workload measures (6) (7) have been used to infer workload states. Physiological methods are based upon measurements of activation or arousal which is a state of preparedness of the body associated with increased activity in the nervous system (7). The question is what these measures do in fact reflect: emotional stress, physical activation, cognitive workload or some unknown combination thereof.

Output result or performance looks at workload as activity or accomplishment i.e. the actual task performance or the quality of task accomplishment. Task performance can be defined as workload in terms of accuracy, timeliness etc. ... and compared to an established task criterion for performance. The problem with these expressions, however, is that they do not always lend themselves to sensitive reflections of workload as an operator can adapt and work harder to achieve equal performance. Performance degradation may occur only after substantial demand and effort increases beyond the range we wish to measure (4) (7).

2. Objective and Subjective Workload Study Methods

A review of the methods available for pilot workload determination indicates that they can be gathered into two complementary groups, i.e. objective and subjective methods.

The first ones use physical or physiological measurements. Workload interpretation is made a posteriori involving subjective appreciations.

The second ones use subjective criteria defined a priori through rating scales or mental load modelling. Whether it is a priori or a posteriori some subjective judgement always seems to play a role in whatever workload evaluations we consider.

The following methods are presently in use or under active development at AIRBUS INDUSTRIE:

Objective Methods

- Static Taskload Method developed by AIRBUS INDUSTRIE.
- Timeline analysis developed essentially in USA by BOEING and McDONNELL DOUGLAS.
- Physiological measures such as Ambulant Monitoring of Heart Rate developed by COCHIN FACULTY and AIRBUS INDUSTRIE.
- Performance criteria measures developed by DUNLAP & ASSOCIATES EAST and AIRBUS INDUSTRIE.

Subjective Methods

- Subjective Assessment of workload such as with the Dynamic Workload Method developed by AIRBUS INDUSTRIE.
- Human Operator Models developed essentially in the USA (SAINT, HOS, PROCURU, ...) and under development in France (MESSAGE) on behalf of AIRBUS INDUSTRIE.

3. Workload and Cockpit Resource Management

The work of a crew is characterized by the appearance of a multitude of tasks or in other words man-machine messages. These seem to arrive simultaneously and to interrupt one another. The processing resources for controlling these information transfer processes are limited and when several processes compete for the same resources eventually there may be deterioration of performance. The process for allocating these resources is called Resource Management. It refers to the way the different task backlogs are prioritized and delegated and to the management of the different human and material adjuncts available to the crew.

Resource management training therefore improves crew coordination, communication, role playing and decision-making skills.

A majority of airline accidents in the last ten years appears to be related to human factor problems and most of these seem to have had as a causal factor some aspects of inadequate resource management. There is a growing awareness and consensus in aviation circles that crews trained in cockpit resource management skills can operate at a higher level of safety and efficiency especially during periods of increased workload (8).

Proper resource management can in fact also serve as a very effective workload control tool.

WORKLOAD ASSESSMENT TECHNIQUES AT AIRBUS

In the following paragraphs are described some of our assessment techniques developed throughout these last years.

The two main methods, the STATIC TASKLOAD ANALYSIS and the DYNAMIC WORKLOAD ANALYSIS are complementary and were used in early 1982 for the first two-person crew certification in the world of a wide body passenger aircraft: the A300 FF, not yet equipped as the A310 with the EFIS, the ECAM or the FMS. Another method was to be added for the A310 certification, i.e. the PERFORMANCE CRITERIA ANALYSIS. The common rationale of all these analytical methods is that they work by full comparison to previously certificated man-machine systems. Since early certification days, all airplanes had been certificated with reference to an existing product. With regard to the cockpit, the Airworthiness Authorities pilots were traditionally basing their judgement and verdict on their own (subjective) assessment in comparison to other cockpits. Without any way to know results in advance, AIRBUS INDUSTRIE effectively made a step forward by developing objective measures and quantifying subjective assessment.

THE STATIC TASKLOAD ANALYSIS**1. Principles of the Methodology**

The static taskload method allows an objective quantitative task analysis of system management procedures that attempts to quantify the ergonomic aspects (visibility, observability, accessibility, operability, monitorability of control and displays) of the man-machine interface of a new aircraft through a direct comparison of procedures with a previously certificated two-person aircraft. It provides quantitative taskload data or in other words objective indications of individual crewmember task demand by measuring the impact of a new cockpit layout, the location and nature of controls and indicators in comparison with a former cockpit layout. After selecting a series of comparable normal, abnormal and emergency procedures each of these procedures is analysed individually for both aircraft. Each task in a given procedure is split into 6 basic actions i.e. look, observe, monitor, reach, operate and monitor (the result of the operation) (9). Each action is linked with a feasibility index which expresses the elementary difficulty to accomplish the action. These visibility, observability, accessibility, operability and monitorability indices are intimately linked with the cockpit layout or hardware. They are expressed in terms of values on a continuous difficulty scale ranging from 0 to 1, the static taskload scale (Figure 2). This scale is adapted from the Cooper-Harper rating scale which is a widely accepted method for subjective assessment of aircraft handling qualities (10).

The laws governing the value of the feasibility indices are determined by means of small mathematical models which were derived from the ergonomic literature and validated through subjective assessments in mock-up by Airworthiness Authority pilots (11) (12).

The method is called static for several reasons.

First, the correspondance between specific actions and feasibility indices is stopped with information processing, problem-solving and decision making activities. These dynamic aspects of mental workload are addressed by the Dynamic Workload Method.

TASKLOAD SCALE	DIFFICULTY	APPRECIATION
0 - 0.2	Very small	Satisfactory
0.2 - 0.4	Small	
0.4 - 0.6	Moderate	
0.6 - 0.7	Significant	Acceptable
0.7 - 0.8	Very significant	
0.8 - 0.9	Severe	Questionable
0.9 - 1.0	Very severe	
> 1.0	Prohibitive	Impossible

Fig.2 Static taskload scale

Second, the method is also based on the assumption that taskload differences between the new aircraft and the reference aircraft principally exist in systems management. The other dynamic workload functions such as flight path control, collision avoidance, navigation, communications and decision making are an integral part of the assessment procedure in the Dynamic Workload Method.

Third, the analysis assumes a strict tasksharing whereby one crewmember (in this case CM1) is flying the aircraft (PF) while the other crewmember (CM2) is mainly involved in operating and monitoring aircraft systems (PNF). In a real world dynamic context tasksharing and task allocation may often be differing event slightly from the prescribed procedures but this can again be taken into account by the other complementary method.

Fourth, the (static) operator returns to his neutral (eye-reference) position after each task and this is also not necessarily the case in a dynamic context.

2. Application to the A300 FF and A310

In the early stages of crew complement research and well before the Presidential Task Force's audit we performed a feasibility study comparing the A310 with the B-737 and the DC-9. The method was then validated in cooperation with French and German Airworthiness Authorities.

For the A300 FF minimum crew certification the Static Taskload Analysis was comparatively applied to the A300 FF and the McDONNELL DOUGLAS DC-9 which has proven in service experience and a reputation for safe and efficient two-man operations (13).

For the A310 minimum crew certification AIRBUS INDUSTRIE benefits from this former experience by comparing to an in-house aircraft, the A300 FF (14).

The Static Taskload Analysis was carried out as follows (Figure 3):

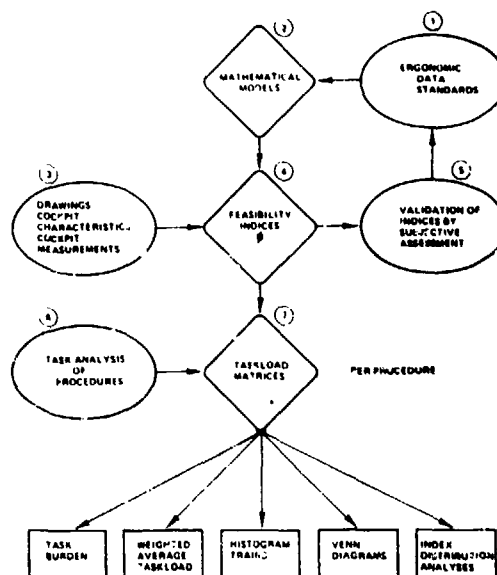


Fig.3 Static taskload analysis flowchart

- (a) Comparable normal, abnormal and emergency procedures were selected for the A300 FF and the DC-9 on one hand, for the A310 and the A300 FF on the other hand; in each case this involved at least 10 normal procedures and 10 abnormal/emergency procedures.
- (b) Task analyses of system management activities were performed for each crewmember of each aircraft with a task breakdown into basic actions (look, observe, monitor, reach, operate, monitor) CM1 (or the left hand seated pilot) being PF, CM2 acting as PNF; the task analyses of the A300 FF and of the A310 were conducted in dedicated mock-ups to aircraft production standard; the task analyses of the B-737 and DC-9 were conducted in a flight simulator with the assistance of a type-rated flight instructor.
- (c) Geometric, time and mechanical measurements from cockpit drawings, mock-up and simulators were used to calculate parameters that are considered in mathematical models of ergonomic feasibility laws.
- (d) Feasibility indices for each action of a task are calculated by means of the mathematical models of each type of action.
- (e) Taskload matrices were compiled for each procedure so that comparisons could be made between the aircraft under evaluation and the reference aircraft, initial results for each crewmember (CM1 or CM2) and for each procedures are expressed in terms of *Burden* and *Weighted Average Taskload*; Burden gives a measurement of the overall amount of work demanded for executing a particular procedure, whereas *Weighted Average Taskload* gives an idea of the average degree of difficulty generated by the execution of a procedure.
- (f) Histogrammic plots of Burden and weighted average taskload for each crewmember were drawn allowing to take first-hand conclusions. Figure 4 illustrates results of the initial studies with respective examples for some normal procedures comparing the A310 and the B-737 and some emergency procedures comparing the A310 and the DC-9.

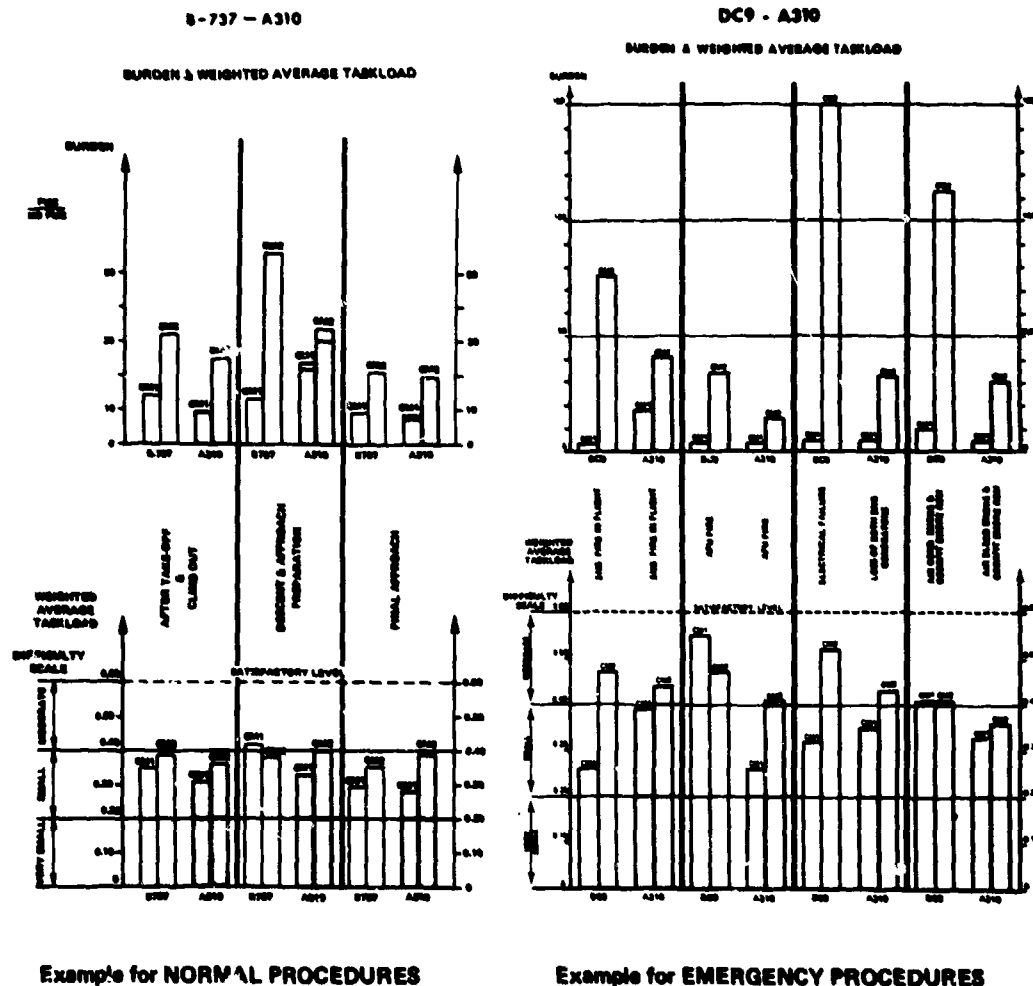


Fig.4 Histogrammic plots of static taskload analysis

3. Discussion of the Results

It appears from these graphic plots that the results of each aircraft under certification were generally indicating decreased taskload burden for each crewmember when compared to their reference aircraft. Burden figures for CM2 are always much higher than for CM1 as the former is carrying out the bulk of the system management work. With regard to the weighted average taskload the individual crewmember figures for the aircraft under certification were generally equivalent to their reference aircraft. More important, however, was the fact that they stay well inside the satisfactory range of the static taskload scale. It is concluded that there are less tasks on the new aircraft and that they are easy to execute.

Several other ways exist to graphically represent the results of the Static Taskload Analysis one of which being Normalized Principal Components Analysis of the taskload matrices (15).

The objective of normalized principal components analysis is to provide a synthetic representation of the information contained in a matrix of p continuous variables and n observations.

The structure of the information included in this matrix would be visible if it were possible to represent the shape of the cloud formed by the n observation points in the p dimensional space of variables. This is not possible when $p > 3$.

Principal components analysis brings a synthetic solution to this problem at the cost of some marginal loss of information.

In this particular way of representation we used procedure matrices whose observation points corresponded with the burden data for normal, abnormal and emergency procedures of both aircraft to be compared. The variable corresponded with the 6 elementary activities in a task. Differentiation of the two aircraft to be compared (the DC-9 and the A300 FF) was done by attributing different codes to the projected observation points. Figure 5 projections for CM1 and CM2 allows to appreciate the relations between points as for example the subcloud of one aircraft may extend beyond or stay within the subcloud of the other aircraft. One can also get an idea of the homogeneity of procedures or of the homogeneity of action burden data associated with the procedures whether the subclouds are clustered or dispersed. In essence this method indicated that as a whole the elementary activities (look, observe, monitor, reach, operate, monitor) on the A300 FF are more homogeneously grouped and centered and therefore less demanding than on the DC-9.

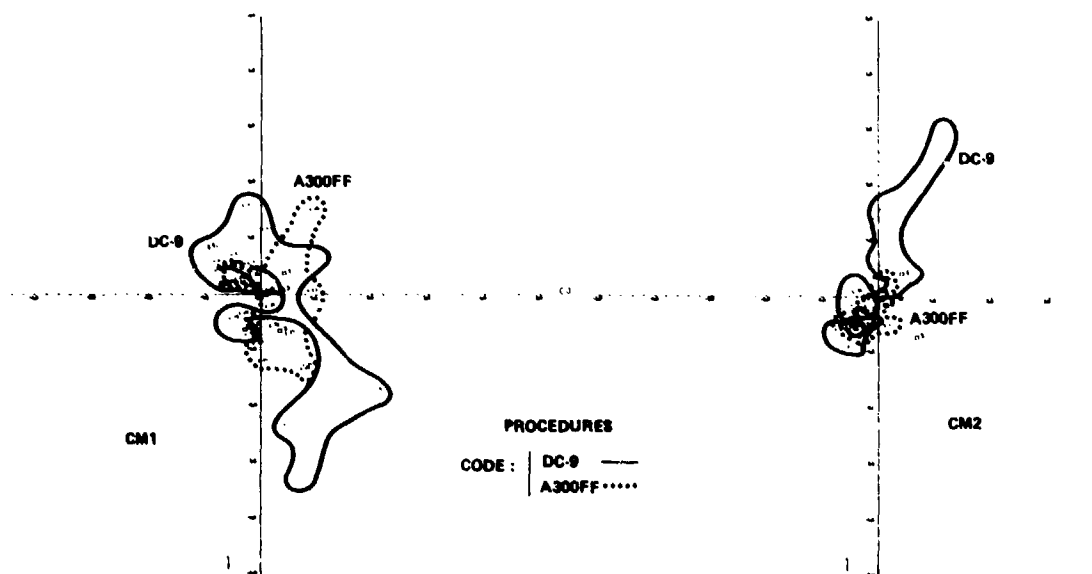


Fig.5 DC-9—A300FF: normalized principal components analysis. All procedures cumulated

Another analytical evaluation of feasibility indices consisted in considering cumulative percentages of actions with feasibility indices on the same 11 intervals between 0 and 1. Separate analyses were conducted for normal and abnormal/emergency procedures amalgamated so as to compare the distributions of specific actions for the A310 to those of the A300 FF. The Kolmogorov-Smirnov two sample test (16) was used to determine if a significant difference existed between the distributions of both aircraft and the direction of any difference detected, i.e. which aircraft was better. The majority of measures showed no statistically significant difference between the distributions of the A310 and of the A300 FF. This is not surprising given the strong similarities between the two aircraft. However, the statistically significant differences which do exist strongly favor the A310 particularly with respect to abnormal/emergency procedures where the ECAM is most instrumental.

In general the Static Taskload Analysis showed that taskload data of the aircraft under evaluation for certification were within or close to the envelope defined by the reference aircraft which by itself already indicated the plausibility of acceptable two-man operations on the new aircraft. Besides this the Static Taskload Analysis also allowed first hand tasksharing

evaluations to be made in mock-up task analyses with early sets of procedures not yet subjected to flight experience. This caused even a major redesign of the electrical system after task analyses for the Avionics Smoke procedure on the A300 FF, suggesting that the method may also be a helpful technique during initial cockpit design.

PERFORMANCE CRITERIA METHOD

The Performance Criteria Analyses presented in the following are a complement to the Airbus Industrie man-machine interface studies originally based on the three functionally related attributes of input load (taskload), operator effort (workload) and output result (performance). They were developed in the aftermath of the US Presidential Task Force recommending increased focus on man-machine interface analysis. In this part two studies are presented performed under contract with DUNLAP & ASSOCIATES EAST (Hartford, Connecticut, USA) to investigate the impact of new digital equipment that was to be installed in the A310 (17).

(A) EFIS Performance Criteria Analysis

1 Principles of the experiment

In March 1982 AIRBUS INDUSTRIE conducted an extensive experiment to determine relative system performance of the new Electronic Flight Instruments (EFIS) versus the conventional electromechanical instruments. The Airbus Industrie's research and development A300 constituted an ideal experimental platform for such a study as it was equipped with the conventional instruments in front of the left pilot seat and with the EFIS configuration in front of the right pilot seat. The aircraft was also equipped with a sophisticated data recorder which can collect most relevant performance measures and record them on magnetic tape for subsequent analysis. The experiment consisted in measuring relative pilot/aircraft system performance in the execution of a specified and relatively demanding circuit (Figure 6) to be flown from each seat. In order to provide for experimental control, a factorial experimental design was employed in which factorial, pilot and instrument/seat were the major variables. Three conditions were chosen to provide a range of situations under which the instruments would be compared and to vary workload for statistical comparison.

These conditions were:

Flight Director:

Flight director and autothrottle system on.

ILS:

Flight director and autothrottle off, but "raw" ILS glide and localizer information available.

NDB:

Flight director, autothrottle and ILS off resulting in a totally non-precision configuration.

Go around was initiated at 100 feet radio altitude for the FD and ILS conditions and at 300 feet for NDB approaches.

To ensure generalizability of the results, two pilots flew each condition from each seat. In order to control for learning and fatigue, each of the twelve situations (2 pilots x 2 seats x 3 conditions) was repeated once in a counterbalanced fashion so that the total design called for 24 trials (12 situations x 2 replications). The flying pilot always wore a helmet-mounted hood to restrict his view to only those instruments on his side of the cockpit. Twelve segments were defined in each circuit (Figure 6) so as to be able to compare the two sets of instruments during individual, homogeneous portions or segments of each circuit.

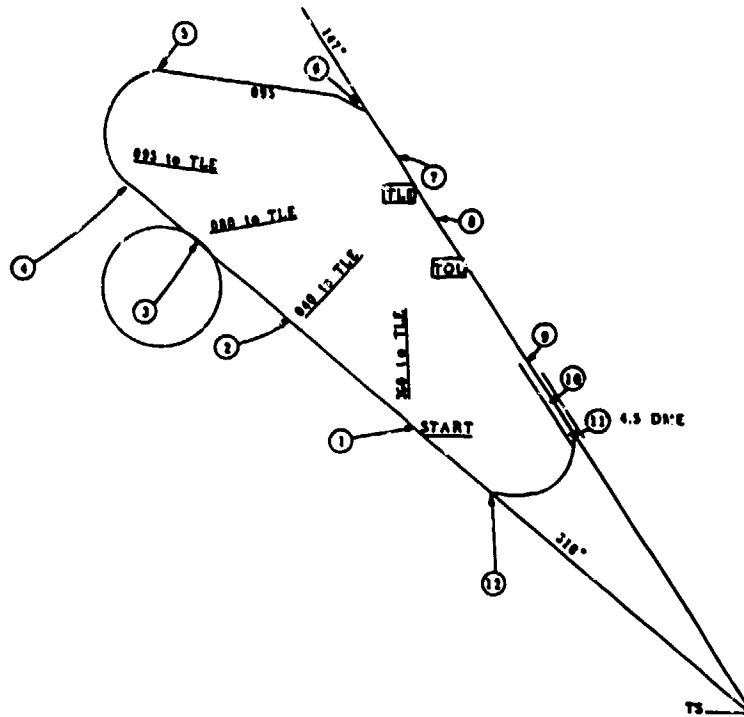
Subjective ratings using the 10-point interruption scale described in the section on Dynamic Methods, were also collected at various points to compare workload levels in either condition.

Four basic measures were calculated for each segment and for many of the 61 parameters recorded:

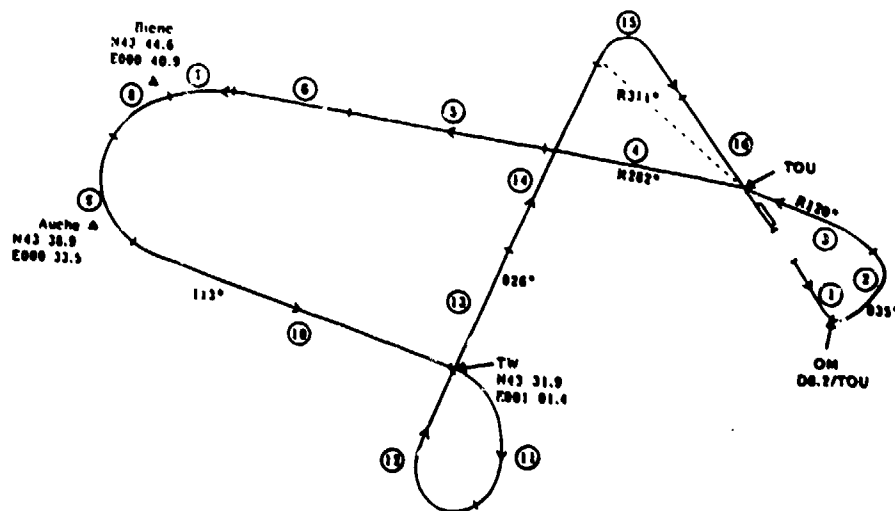
- **Mean:**
The numerical time-based average of the parameter.
- **Standard Deviation:**
A standard measure of the amount of variation around the mean, the standard deviation has proved to be an excellent measure of system smoothness and stability.
- **Transitions Through Zero:**
The number of sign changes per minute for those parameters which can have both positive and negative values, this rate also measures stability and the extent of control inputs needed to achieve the observed mean and standard deviation.
- **Reversal Rate:**
The number of direction reversals per minute of controls and control surfaces, reversal rate is a direct measure of the control activity and the taskload of the pilots.

The major analytical technique chosen for the instrument comparison was multi-dimensional analysis of variance (ANOVA) (18).

This technique separates the variation in a dependant variable (the various measures derived from the flight parameters) into the components which can be attributed to each of the independent factors (in this case pilot, conditions, seat, segment and replication) or interactions between or among the factors, and a component (error) which cannot be attributed to any of the factors or interactions. The amount of variation associated with a factor by the ANOVA calculation procedure can thereafter be tested for statistical significance.



EFIS Study Circuit



FMS Study Circuit

Figure 6

2 Presentation of the results

The statistical results of this method clearly show differences as a function of the seat from which the aircraft was flown. In all 2310 statistical tests of significance were examined (14 effects per measure x 165 measures retained for ANOVA) exclusive of the replication factor. It was found that the number of significance effects involving the seat factor was well above what would likely be produced by chance (116 when testing at the 0.05 level of significance meaning less than 5 chances out of 100 that the effects were produced by chance).

Given that the existence of a specific difference has been shown, it could have favored the EFIS or the conventional instruments. However, all of those measures which could be interpreted in terms of smoothness or precision of flight favored the EFIS. The remaining differences either favored the EFIS or showed variation across the experimental factors which could not reasonably be interpreted as favoring either instrument type.

These measures were for example :

- Pitch speed or rate changes through zero were significantly higher for the EFIS (18.46 versus 17.48 per minutes for conventional) across all flight conditions and in each one individually. At the same time, the standard deviation of pitch rate, a measure of smoothness, was the same or lower for the EFIS. Together, these measures showed that the pilots were making finer corrections around the criterion value (higher rate through zero) while still maintaining or improving overall smoothness.
- Elevator position reversal rate was also significantly higher in all conditions when the aircraft was flown with the EFIS (32.15 versus 30.12 per minute for conventional). Also the standard deviation of elevator position was not significantly different for both types of instruments.

Thus greater precision was accomplished with the EFIS with equivalent smoothness.

- Engine 1 power lever angle reversal rate was significantly higher for the EFIS flown trials (21.49 versus 20.37 per minute for conventional). The difference was most pronounced during the ILS and NDB conditions in which the autothrottle system was disengaged and the pilots had to manage the engines manually to track the target airspeeds.

It is interesting to note that the superior performance of the EFIS was particularly pronounced for those measures which are related to information which is displayed in a new, more precise fashion on the electronic instruments. The smoother and more accurate performance of the pilot/aircraft system may be simply because pilots tried harder when flying with the EFIS or because the EFIS presented more or better information for flight maneuvering. The results must be interpreted with the understanding that neither pilot had extensive experience with the EFIS. It is reasonable to hypothesize that a greater level of pilot familiarity with EFIS would have shown an even larger performance benefit when flying in the normal to moderately difficult flying situations experienced during the experiment.

The increased reversal rates of elevator position and engine power levers and the higher rates through zero of pitch speed are indications of increased taskload since all flying was under manual control of the pilots. With regard to workload the rating method described in Part II on Dynamic Methods shows a slightly lower mean workload for the EFIS seat but this difference was however not statistically significant. Since the experiment was not designed to examine this aspect specifically workload in both configuration is concluded to be equivalent, greater precision being accomplished with the EFIS.

(B) FMS PERFORMANCE CRITERIA ANALYSIS

A similar experiment was carried out by Airbus Industrie in January 1983, to assess the performance of the A310 aircraft/pilot system with and without the use of the FMS when the autopilot was engaged and to examine the ability of pilots to use the FMS information directly without aid of the autopilot.

The experiment consisted in measuring relative pilot/aircraft system performance in the execution of a specified and relatively demanding circuit which combined a SID and STAR (Figure 6).

Three experimental conditions were studied :

- NAV :
"Normal" flying with the Flight Management System commanding the autopilot and the pilot monitoring horizontal track and entering altitude and speed adjustments.
- STANDARD :
"Normal" flying without the Flight Management System in which the pilot commanded all course, altitude and speed changes through the autopilot.
- MANUAL :
Manual flying without intervention from the autopilot or help of the flight director but using FMC track information as displayed on the ND for navigation.

As in the EFIS study, to ensure generalizability of results and increase the precision of analyses, it was desirable for each pilot to fly each condition twice in a counterbalanced design. However, to permit data collection to be undertaken in a single flight, the NAV condition was only flown once by each pilot. In addition subjective workload ratings using the 10-points interruption scale were again collected as in (A) to compare levels in either condition.

Similarly to the preceding study, six basic measures were extracted for each segment and for many of the 75 parameters recorded. The major analytical technique obviously was again multidimensional analysis of variance (ANCOVA) (18).

A statistical examination of the significant effects in this experiment shows three clear patterns of findings.

First, many of the significant differences appear to relate to whether or not the autopilot was engaged. The autopilot clearly has vastly different response characteristics to those of the pilots. In general, the autopilot allows "error" to build up more before it responds than do the pilots.

It then brings the system back to a nominal state with little overshoot or additional correction. Since the autopilot was engaged in both the NAV and STANDARD conditions, this "autopilot" effect causes them to appear quite similar and quite different from the MANUAL condition.

The second pattern of results, relates to the similarity between the NAV and MANUAL conditions on certain parameters which relate to the way the aircraft maneuvers in the horizontal plane. It would appear that the increased precision of track specification by the FMC, whether manually or by the autopilot, results in better flying performance on these parameters.

The third clear pattern in the results highlights superior performance when flying in the NAV condition.

These findings suggest differences in the smoothness of the tracks flown during the NAV trials when compared with the other two conditions. In particular, the extremely low yaw rate with an associated low standard deviation of rudder position point to significantly less stressful and more comfortable flying with the FMS.

As commented in the section on Dynamic Methods the expected ordering of conditions with respect to workload was achieved, NAV showing lowest, MANUAL highest. The trend displayed here is nonetheless made even more noteworthy by the fact that neither pilot had extensive experience with the FMS and therefore could have been expected to show some degree of extra preoccupation with flying in these conditions.

CONCLUSION

The practice of man-machine interface analysis clearly got an added impetus with the approach to the issue discussed in this paper. It culminated into a battery of methods that not only dealt with that particular aspect but much more generally investigated the impact of new technology and its match with the crew and operations. Man-machine interaction analysis precisely is in the business of examining these matches. Rather than insisting excessively on workload it concentrates on information-transfer which, we believe, is the essential parameter of the interface equation.

The common rationale of all our methods is that they work by comparison to previously certificated man-machine systems. They were launched as an important and risk-taking validation exercise several years ago without any prior certainty as to what they would produce as results. Clearly, the practice of flight testing by cross-reference to former designs is classical and justified but there is an upcoming need to develop integrated workload and performance standards which would potentially alleviate or delete this requirement for comparison.

A step in this direction was performed by validating some of the previously mentioned work with regard to Dynamic Workload Analysis (see Part II). It was done by means of Performance Criteria Analysis, described in the preceding paragraphs and Ambulant Monitoring of Heart Rate, mentioned in Part II.

PART II

DYNAMIC METHODS

1. PRINCIPLES OF THE METHODOLOGY

The Dynamic Workload Method is a subjective, qualitative technique to assess the workload resulting from the interaction of all piloting and management functions mentioned in FAR25, Appendix D. It addresses mental effort due to the time pressure, information processing and emotional stress whilst piloting the aircraft under a variety of normal, abnormal or emergency conditions. With this method the point is made that the mental effort associated with collecting and processing information and the making of decisions is much more prominent in pilot workload than the actual physical implementation of decisions through actuation of controls. Hence covert efforts which occur in the planning, monitoring and decision-making processes are of particular importance.

The method's application basically consists in a concurrent assessment of workload by the pilots and by an observer-pilot. This is done by means of a common workload scale modelled after the Cooper-Harper scale. The first scheme adopted at AIRBUS INDUSTRIE consisted of a 5-point scale for pilots and an overlapping 7-point scale for observers. This dual workload scale was used for the A300 FF workload campaigns and consisted of one rating choice for each workload category for the pilots.

Observers disposed however of two rating choices for both the low and moderate workload categories (19) (20). The experience of the A300 FF certification showed that the rating activity was quite unintrusive to pilots, that low workload did not need two categories but that heavy load deserved a choice selection.

WORKLOAD ASSESSMENT		CRITERIA			APPRECIATION
		RESERVE CAPACITY	INTERRUPTIONS	EFFORT OR STRESS	
LIGHT	2	AMPLE	—	—	VERY ACCEPTABLE
MODERATE	3	ADEQUATE	SOME	—	WELL ACCEPTABLE
FAIR	4	SUFFICIENT	RECURRING	NOT UNDUE	ACCEPTABLE
HIGH	5	REDUCED	REPETITIVE	MARKED	HIGH BUT ACCEPTABLE
HEAVY	6	LITTLE	FREQUENT	SIGNIFICANT	JUST ACCEPTABLE
EXTREME	7	NONE	CONTINUOUS	ACUTE	NOT ACCEPTABLE CONTINUOUSLY
SUPREME	8	IMPAIRMENT	IMPAIRMENT	IMPAIRMENT	NOT ACCEPTABLE INSTANTANEOUSLY

Fig.7 Dynamic workload scale (A310)

The common A310 workload scale therefore consisted of 7 points from 2 to 8 (Figure 7) which offers one rating choice for the low workload category 2, two rating possibilities for the moderate 3, 4 and the high workload 5, 6 categories. The two remaining rating alternatives concern extreme 7 and supreme 8 workload cases that impose strict scrutiny during the post-flight analyses (21) (22). A description of the scale by means of selection criteria is provided for initial guidance but due to the diversity of workload connotations across individuals (23) AIRBUS INDUSTRIE insists that pilots and observers be strictly guided by their own personal interpretation of the term workload. The objective of the common pilot and observer scale is to give pilots and observers a frame of reference on workload acceptability without imposing a definition or a viewpoint. Pilots and observers are however asked to rate workload and not taskload nor performance.

Practically, flight crews are subject to a comprehensive set of both simulator and real flight scenarios throughout which they are required to provide frequent and prompt ratings of workload. Pilots are trained to rate their workload experience upon the observer's request.

The observer-rater first introduces his assessment of estimated pilot workload by means of a rating box (Figure 8), and thereby triggers the corresponding pilot's green cue light. This effectively requests the pilot's response which is implemented through activation of the appropriate push-button on the pilot rating box which is on the glareshield.

The basic instruction is for the observers to request and provide a rating whenever they feel workload since the last rating has changed or in the absence of such variation if a substantial amount of time (more than 5 minutes) has elapsed since the previous rating. The rating system is designed such that its operation imposes minimal interference, although pilot response time to a rating request may be significant as pilots are instructed to give priority to their immediate work. It should be noted

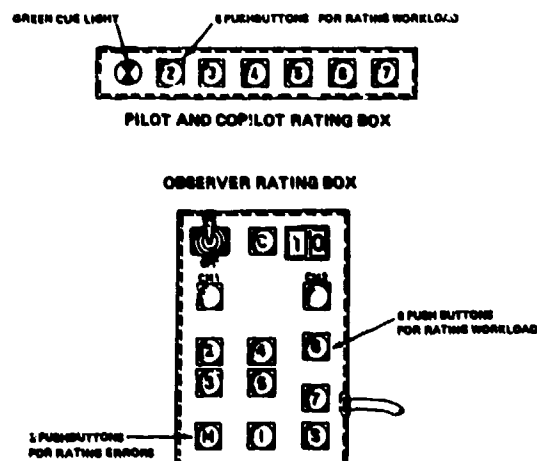


Fig.8 Workload rating boxes

that observer-raters are also asked to keep track of errors by rating them according to their gravity as proposed by a 3-point scale. These performance measures can also be introduced on the more elaborated observer pushbutton box. The three categories of errors which are considered are as follows:

- Minor errors (M): slip or error fixed promptly.
- Important errors (I): important errors, corrected or uncorrected anomalies or safety-unrelated errors uncorrected.
- Safety-related errors (S): errors that may affect safety in the longer term whether corrected or not.

The rationale behind the coupled pilot-observer rating procedure is that each pilot is intimately facing his workload situation and that this close implication may sometimes bias his appreciation either towards an overstatement or towards an understatement. Similarly, the observer is following at the same time the corresponding pilot's workload situation but this independent relation may also bias his appreciation either towards an overstatement or towards an understatement. AIRBUS INDUSTRIE's basic hypothesis is that the true picture may in fact be lying between both appreciations of workload. In the first case too much or too little emphasis may indeed be put on covert processes, in the second case too much or too little emphasis may be put on overt behaviour. Close attention is therefore given to the degree of concordance of pilot and observer opinions and this is the subject of further analysis in the discussion of results in order to validate the overall method.

In the A300 FF minimum crew certification the Dynamic Workload Analysis was applied during both a simulator campaign and a real flight campaign (Figure 9). The simulator campaign involved three programmes, one on the A300 FF and two on reference aircraft i.e. the B-737 and the DC-9 (19) (20). This procedure of comparing with well-established two crew aircraft is similar to the approach of the Static Taskload Analysis. The objectives were to provide the Airworthiness Authority observers with baseline references on two-man aircraft which have proven in service experience and to calibrate their use of the workload scale by means of common scenarios. This was considered essential prior to their participation to the A300 FF simulator and flight programmes. The participating crews comprised one complete Airworthiness Authority crew and two crews consisting of a captain (CM1) from the Airworthiness Authorities assisted by a pilot (CM2) from AIRBUS INDUSTRIE.

In the A310 minimum crew certification the Dynamic Workload Analysis was also applied during both a simulator and a flight campaign (Figure 10). For both campaigns AIRBUS INDUSTRIE was able to benefit from the former experience by comparing to an in-house aircraft, the A300 FF, as in the Static Taskload Analysis (13) (14). The extent of the A310 minimum crew certification exercise was however much larger than with the A300 FF because up to 7 different crews participated to both the simulator and flight campaigns. These consisted of one complete Airworthiness Authority crew, two crews with a captain (CM1) from the Airworthiness Authorities assisted by a pilot (CM2) from AIRBUS INDUSTRIE and four crews each with a captain (CM1) from one of the 4 launching airlines also assisted by a pilot (CM2) of AIRBUS INDUSTRIE.

The preparation of the various flight scenarios was largely inspired by such training techniques as L.O.F.T. (line orientated flight training) and combined operational difficulties with in-flight technical problems involving abnormal and emergency situations.

A brief examination of cumulated results helps to appreciate the procedure adopted for comparing the results of the aircraft under certification with the reference aircraft. The principal numerical tools used for this were the cumulated rating distributions presented under histogram form (see Figure 10). These histograms permit to get an idea of the frequency of timewise distribution of ratings throughout any particular campaign. In particular the addition of the low to moderate workload categories shows that the A300 FF workload levels are at worst equivalent but generally even better than those of the B-737 and the DC-9 when considering common scenarios exercised on the simulator.

DYNAMIC WORKLOAD ANALYSIS A300FF	DYNAMIC WORKLOAD ANALYSIS A310	
3 CREWS + 4 OBSERVERS EUROPEAN NETWORK 16 SCENARIOS	7 CREWS + 8 OBSERVERS EUROPEAN NETWORK + LONG DISTANCE FLIGHTS 10 SCENARIOS	
	MINIMUM CREW CERTIFICATION	AIRLINE ORIENTED ANALYSIS
. VALIDATION OF METHOD IN A300 SIMULATOR 44 H . WORKLOAD EVALUATION ON B-737 12 H . WORKLOAD EVALUATION ON DC-3 12 H . CREW TRAINING ON A300FF SIMULATOR 30 H AIRCRAFT 8 H . WORKLOAD EVALUATION IN SIMULATOR CAMPAIGN ON A300FF 48 H (36 FLIGHTS) . WORKLOAD EVALUATION IN FLIGHT CAMPAIGN ON A300FF 68 H (56 FLIGHTS) . CAT II WORKLOAD EVALUATION IN A300FF SIMULATOR 4 H	. 3 CREWS WITH AA/PI PILOTS . TRAINING TO METHOD IN A300FF SIMULATOR 14 H . CREW TRAINING ON A310 SIMULATOR 48 H AIRCRAFT 8 H . WORKLOAD EVALUATION IN SIMULATOR CAMPAIGN ON A310 48 H (42 FLIGHTS) . WORKLOAD EVALUATION IN FLIGHT CAMPAIGN ON A310 67 H (43 FLIGHTS)	. 4 CREWS WITH AL/PI PILOTS . TRAINING TO METHOD IN B-737/DC-8/BAC1-11 SIMULATOR 16 H . CREW TRAINING ON A310 SIMULATOR 31 H AIRCRAFT 8 H . WORKLOAD EVALUATION IN SIMULATOR CAMPAIGN ON A310 16 H (12 FLIGHTS) . WORKLOAD EVALUATION IN FLIGHT CAMPAIGN ON A310 27 H (24 FLIGHTS)

Fig.9 Dynamic workload analysis programmes for A300FF and A310

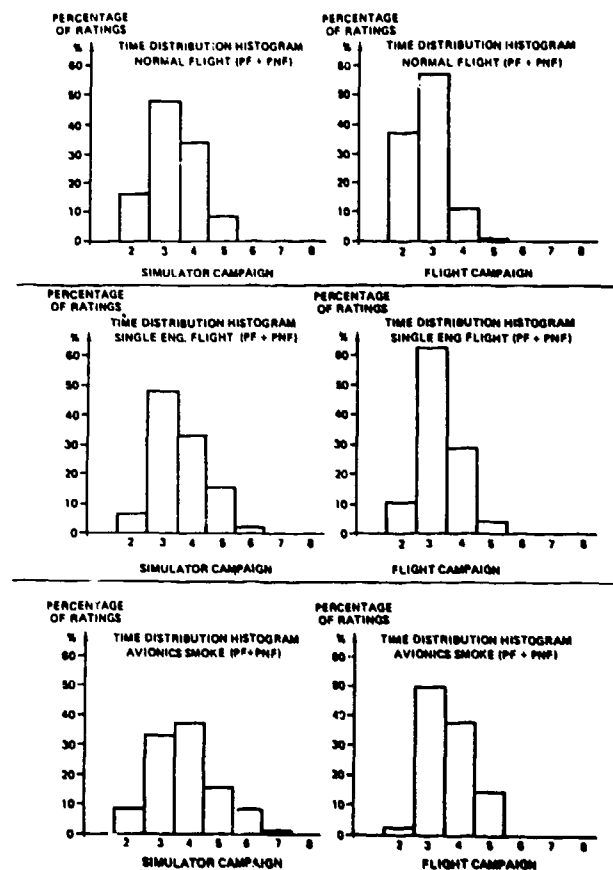


Fig.10 Histograms of workload ratings

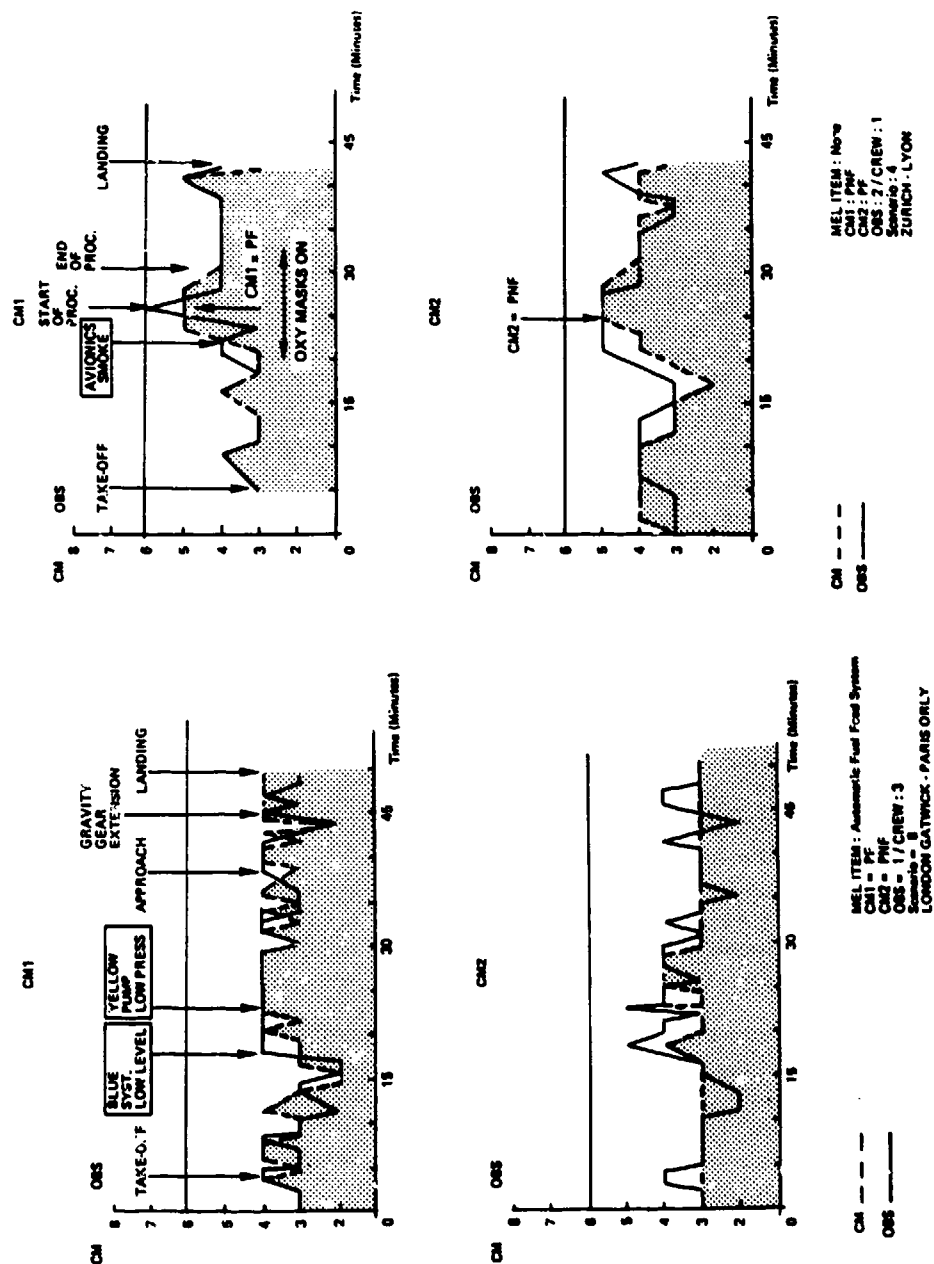


Fig.11 Workload timelines

It was also demonstrated that the various workload categories are at least similarly occupied on the A310 as on the A300 FF so that it was established that workload levels on the A310 remain within the envelope defined by the A300 FF for the common scenarios of both the simulator and flight campaigns (Figure 10).

The application of this methodology for minimum crew certification with the A300 FF and the A310 in both the simulator and real flight campaign is described in more detail elsewhere (24).

2. DISCUSSION OF THE METHOD

The actual results of workload ratings as provided by pilots and observers may be presented along a timeline. In order to follow the graphic timelines more easily, main events were plotted so as to link the workload spectra with the scenarios being exercised. Workload ratings are joined by solid lines for the observer and by interrupted lines for the pilot (Figure 11) (19) (20) (21) (22). This helps to visually appreciate the concordance between pilot and observer ratings and its evolution through time. It should be recalled for the appreciation of this that during the A300 FF certification pilots had a 5-point scale from A to E and observers a 7-point scale from 2 to 8. Moreover observer ratings 2 and 3 corresponded with rating A for low and observer ratings 4 and 5 with rating B for moderate workload. The experience with the A300 FF indicated that for the A310 certification only one rating choice should be given for the low workload category 2, whereas two rating possibilities should be provided for the moderate 3 and 4 and high 5 and 6 workload categories respectively.

Before reaching for conclusions it was judged necessary to validate the Dynamic Workload Methods with some elementary statistical tests as is commonly done in marketing and social research. The first tests had to prove the good concordance of pilot and observer ratings in order to validate the principle of the coupled rating procedure. An adequate quality index needs to take into account the values of pilot/observer differences of opinion as well as the amount of time this was sustained throughout the flight. It was moreover hypothesized that the workload scales could be considered as continuous numerical scales with constant intervals between the workload levels. The selected quality index expressed as divergence index integrates throughout the whole flight the areas of the graphic timelines where pilot and observer have made different workload assessments and refers this to the whole envelope area generated by the observer's graphic timeline. This index formulation is in fact truly reflective of the reader's appreciation of the comparison of pilot and observer graphic timelines.

The results brought to our attention that :

- (a) Better agreement between pilots and observers was reached during the flight campaign than during the simulator campaign both for the A300 FF and A310 minimum crew demonstrations (12 simulations out of 35 for the A300 FF, 40 simulations out of 54 for the A310, 28 flights out of 50 for the A300 FF and 45 flights out of 60 for the A310 with a divergence coefficient smaller than 3.33%).
- (b) Biggest divergence indices were recorded during the crew's first flights in each campaign which together with (a) seems to indicate that there is an adaptation to the rating activity.
- (c) Overall a relatively good overall concordance of opinion was obtained especially between observers and flying pilots ; concordances were even improved on the A310 versus the A300 FF after adoption of the adapted common pilot-observer workload scale.
- (d) Divergence indices were relatively constant throughout a flight's history i.e. when they started low or high they remained low or high throughout a flight.
- (e) Some crews and observers were better in reaching agreement than others but overall divergence was low.
- (f) Divergence indices appear to be independent of scenario difficulty.

Another very simple check helped to confirm some of these conclusions since it appeared that when converting the ratings collected on the A310 to the scale adopted on the A300 FF :

- (a) Full agreement between pilot and observer was reached :
 - For 68.1% of ratings on the A300 FF but for as much as 76.7% of ratings on the A310 during the simulator campaigns,
 - For 80.8% of ratings on the A300 FF but for as much as 84.8% of ratings on the A310 during the flight campaigns,
- (b) Disagreement by just one workload category between pilot and observer was reached :
 - For 31.0% of ratings on the A300 FF but for only 22.85% of ratings on the A310 during the simulator campaigns,
 - For 18.85% of ratings on the A300 FF but for 15.6% of ratings on the A310 during the flight campaigns,
- (c) Disagreement by more than one workload category between pilot and observer was reached :
 - For 0.7% of ratings on the A300 FF but for 0.45% of ratings on the A310 during the simulator campaigns,
 - For 0.35% of ratings on the A300 FF but for 0% of ratings on the A310 during the flight campaigns.

The final verification looked at the possible relation of errors with workload. Statistical work on these data showed that :

- (a) The classical shaped curve (4) between performance and workload could not be verified indicating that pilots never go to situations where they could but make errors.
- (b) The simulator campaigns brought proportionally more errors than the flight campaigns possibly because simulator scenarios were somewhat harder or were more difficult to execute.
- (c) There was no direct relationship between scenario difficulty and errors.

3. EXAMPLES OF WORKLOAD RATING APPLICATIONS

The Performance Criteria Analysis presented in Part I of this chapter respectively concern man-machine interface experiments on EFIS and FMS (25) (26).

The EFIS-experiment involved flying a 15-minute take-off, circuit, approach and landing/go-around task to compare the system with conventional instruments.

This involved three configurations (flight director ILS, raw data ILS and NDB non-precision) by two pilots with two replications on each system.

A new subjective rating scale based on the concept of interruption or bother was used so as to enable workload comparisons. In this 10-point scale, the "1" corresponds to little bother at an optimal time in the flight for the pilot to be interrupted, while the "10" corresponds to a big bother at an inopportune time for interruption. Ratings were requested at 12 predetermined points in the segments of each circuit.

In addition, the Flight Test Engineer was requested to ask for additional ratings at his own discretion.

An examination of the table below shows a slightly lower mean workload for the EFIS seat, but this difference was not statistically significant. The EFIS also exhibits a slightly larger dispersion that is most probably due to the relative inexperience of both pilots with this equipment when compared to the conventional instruments. The rating scores by condition showed that the pilots rated the Flight Director trials as having lowest workload, followed by the ILS trials with the NDB trials rated as highest in workload. This effect was significant when tested by a one-way analysis of variance and indicates that the expected ordering of conditions with respect to workload was, in fact, achieved.

The results showed that the EFIS were not associated with any higher workload event with pilots who were relatively inexperienced in the use of the new electronic instruments.

Summary of subjective rating

	Mean	Standard Deviation
Left seat	6.7	2.3
Right seat	6.5	2.5
Flight Director	5.4	1.9
ILS	6.9	2.1
NDB	7.7	2.6

The FMS-experiment similar to the preceding involved flying a 25-minute take-off, SID, STAR and landing/go-around task to compare normal flying with the FMS commanding the autopilot (NAV), first with normal flying without the FMS but with autopilot (STD) and second with manual flying without autopilot or flight director but with FMS for navigation (MAN). Obviously the same "bother" scale was used in this experiment when requesting ratings from two pilots in the NAV (no replication), STD and MAN condition (two replications each).

The distribution of workload ratings shown in the table below shows that the actual mean values are in the predicted direction with the NAV condition showing the lowest value, followed by the STANDARD condition and the MANUAL condition being the highest. However, the differences in the mean values were not sufficient to yield significance with only 90 ratings.

Summary of subjective rating

	Mean	Standard Deviation
NAV	5.5	2.0
STANDARD	6.1	1.6
MANUAL	6.3	1.8

The trend displayed here is nonetheless made even more noteworthy by the fact that neither pilot had extensive experience with the EFIS or FMS and therefore could have been expected to show some degree of extra preoccupation with flying in these conditions.

As a conclusion the performance gains observed for both the EFIS and FMS (reported in Part I) were not associated with any increase in the workload perceived by the pilots in the experiments. Flying with the EFIS is rated as bringing lower workload than with conventional equipment, using the FMS is associated with lower workload than trials flown without it. Although neither of these differences were statistically significant, the results provided the clear implication that pilot workload would be positively influenced by the introduction of these new electronic flight systems.

WORKLOAD MODELING DEVELOPMENTS

1. Workload Index Development

Today's certification process involves a lengthy set of test flights during which pilots give subjective workload ratings. It was reasoned that the entire process would be greatly simplified and made more objective and precise if a model relating workload to system performance measures (described in Part I of this chapter) were to be generated and validated. The flights for the EFIS — instrument comparison constituted an ideal setting on which to superimpose this research conducted in cooperation with DUNLAP & ASSOCIATES EAST.

For this effort the idea was to develop a statistically appropriate mathematical model relating the subjective workload ratings to objectively measured flight parameters. Thus, it was clear that the times at which parameter values were important were those times at which ratings were requested. This precluded the need for any other breakdown of the circuit and defined 402 points in time for which values had to be derived from the data tapes.

The dependent measure for the experiment was the rating given on the 10 point numeric scale described earlier.

The independent measures from which an attempt would be made to predict ratings had to be constructed from the 61 parameters mentioned earlier.

Multiple regression with ratings as the dependent variable was chosen as the analytical tool for building a workload index. Multiple regression is a generalized statistical technique which predicts a dependent variable using one or more independent variables.

Only those independent variables which represented input by the pilot or the response of the aircraft system were considered for inclusion. It was decided that the instantaneous value of these variables at the time of the rating would generally not be an appropriate measure for several reasons. First, the single second value at the time of a rating might be a transient and not truly representative of the parameter being observed. Second, while the pilots were asked to give an instantaneous judgment, experience and the literature show that they would tend to base their rating on their impression integrated over some time period. Third, there was a varying and unmeasured response delay between the request for a rating and the flying pilot's response. Hence, choosing only one second's data might introduce needless error.

The situation suggested that some smoothing of data was needed.

For this study, multiple regression was applied in a stepwise fashion. Using this technique, a set of variables is chosen and individual independent variables are allowed to enter the model one by one on the basis of some pre-established statistical criteria. This procedure is generally used when one wishes to isolate a subset of available predictor variables that will yield an optimal equation with as few terms as possible.

A model of this type appears to have the potential, if validated, to result in an excellent and extremely useful tool for measuring workload.

It is important when considering a model such as this to examine the reasonableness or apparent face validity of the measures which the stepwise procedure has brought into the equation. The most important variables relate to the flight director pitch order which is essentially an error measure. The literature tends to indicate that the perception of error is often related to workload. Likewise, several acceleration measures enter the equation as theory would predict. In fact, all of the measures in the model appear reasonable because they are either the direct result of pilot actions, e.g. elevator position reversals; represent error conditions which must be attended to, e.g. flight director orders; or are related to the stability or smoothness of flight, e.g. pitch angle. Thus, it can be concluded that the model likely has physical meaning and is consistent with theory.

In spite of all of these considerations, care must be exercised in using this model as any other until it is validated. Only two pilots took part in the experiment and only a relatively narrow range of workload was examined. No data were collected under extremely low workload conditions, such as cruise, or extremely high, such as associated with an emergency.

2. Ambulatory Monitoring of Heart Rate

Among all physiologic parameters that may objectivate the impact of task performance, heart rate and heart rate variability appear to be very responsive indicators of the activity of the sympatho-adrenergic system and consequently of the adaptation of the human being to physical exercise, to mental load or to a situation of emotional stress (27) (28) (29) (30).

The measurement of the periodicity of electric cardiac activity by means of electrocardiographic recording (ECG) appears to be a most accurate way to study heart rate. A method of ambulatory monitoring of heart rate for transport pilots was developed for this purpose by J.P. FOUILLOT and J. REGNARD of the Laboratoire de Physiologie at Cochin Faculty of Medicine in Paris.

A miniature magnetic tape-recorder, records the ECG, a 60 Hz signal produced by a quartz clock and an identification signal introduced for synchronization purposes by means of an event marker-button. An observer keeps an activity log on a paper grid with the help of an electronic chronometer. In this way flight deck activity is cut in a series of time sequences which are all identified by a four-digit code. This observation of cockpit activity is synchronized with the recording of ECG by means of the event marker at the start of the flight. Cardiac period (RR-interval) is measured by the time elapsed between two QRS-waves detected by means of an analog system. The measurement is made from the clock.

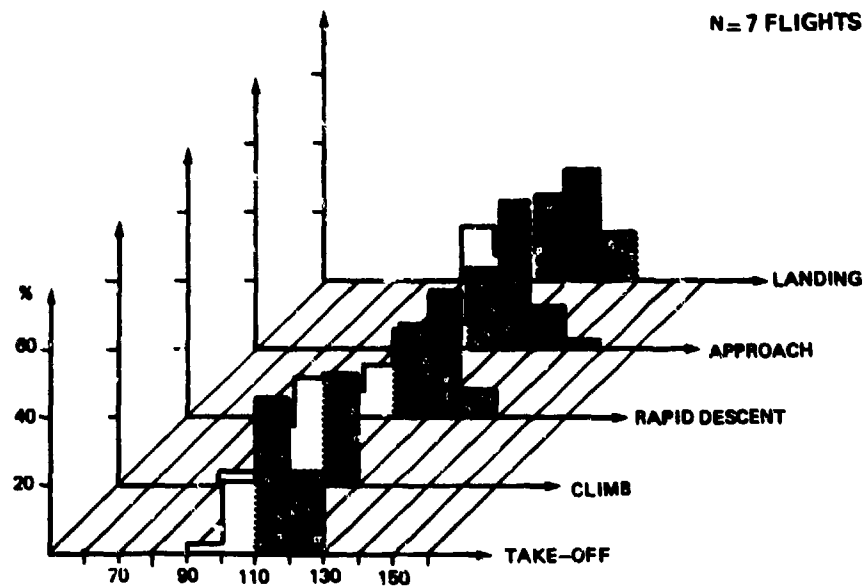
In a first approach these recordings were therefore processed to obtain RR interval histograms for all flight sequences. The representation of heart rate variation by means of histograms for cardiac periods (or RR intervals) is an effective way to condense the abundant information of each flight sequence. These RR interval histograms are presented using 10 classes of heart rate categories expressed in heart beats per minute; each class (from 60 to 69 bpm until from 160 to 169 bpm) corresponds with the percentage of the total number of heart beats detected for that category during the sequence.

The RR interval histograms provide a synthesis of heart rate response corresponding to a flight sequence which is a microscopic view with regard to the whole flight. In order to provide a macroscopic view of heart rate variation, histograms are cumulated per flight phase, per scenario, and per pilot function over the whole population for 7 crews involved in the A310 two-man crew certification.

Scenarios involving non-major failures are not associated with any increase in heart rate. On the opposite scenarios involving a degradation of flight conditions and a rapid change of flight plan such as rapid descent, electrical smoke/fire or single engine flight bring an increase in heart rate indicating the possible occurrence of mental load or emotional stress. Heart rate histograms for rapid descent are shown in Figure 12.

**A310 FOR INTERVAL HISTOGRAMS
FOR PF (RAPID DESCENT)**

N=7 FLIGHTS



**A310 FOR INTERVAL HISTOGRAMS
FOR PNF (RAPID DESCENT)**

N=7 FLIGHTS

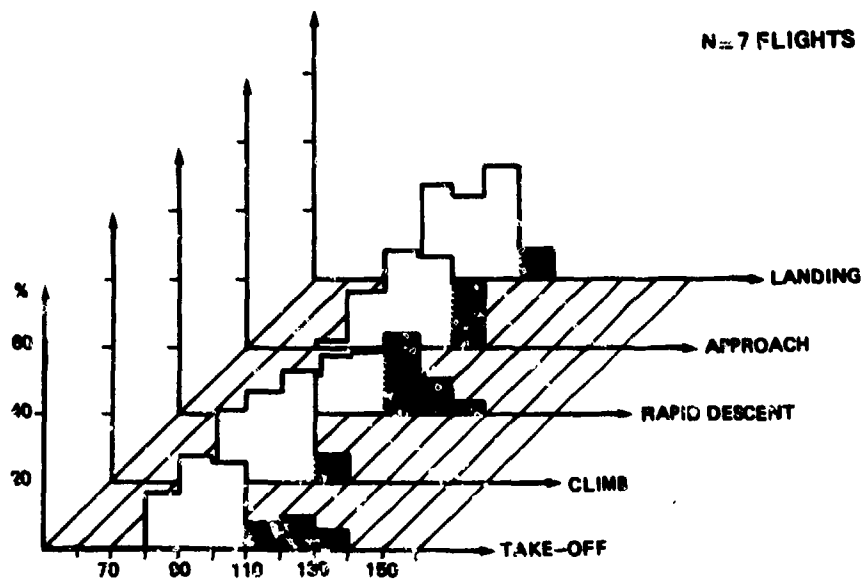


Fig.12 R-R interval histograms A310

For the flying pilot (PF) these histograms demonstrate that there is no increase in heart rate frequency during engine failure recognition and checklist processing by the other pilot. For the non-flying pilot, this sequence results, however, in a heart rate increase as he is handling the check list and applying the procedure. The flying pilot undergoes an increase during approach through landing with a maximum during go around (if any). The non-flying pilot has a decrease in these sequences compared to the sequence corresponding with the failure introduction. These general trends are summarized in the table below which shows heart rate percentages above 109 bpm for the various phases :

	Climb	Failure introduction + C/L	Approach	Go-around	Landing
PF	7%	7.6%	35.4%	56%	35%
PNF	9.9%	25%	12%	20.2%	11.7%

It should be reminded that take-off, go around and landing phases were, however, of relatively short duration lasting no longer than 90 seconds.

In a second approach eight indices of heart rate and heart rate variability were processed at each evaluation of workload rating by either pilot or observer.

We studied the correspondences between these heart rate and heart rate variability indices and pilot ratings simultaneously recorded using the method of factorial analysis of Benzecri (31). This method is applicable to any given table of positive values having rows of N individuals (in this study the flight sequence corresponding to workload rating evaluations) and columns of N variables (in this study heart rate and heart rate variability indices and workload ratings). This permits to represent the sets of variables and individuals in the system of orthogonal basis defined by the factorial axis (32). In order to apply factor analysis of correspondences homogeneous data have been obtained by defining classes within boundaries for all the above mentioned variables.

From the A310 flights' material we have inventorized 3032 sequences and divided each variability index into seven classes going from lower to higher values.

Coding of workload ratings and various variability indices helps to extract 64 modalities.

The initial data table has dimensions of (3032,64) and figure 13 shows the deduced 2 dimensional factorial space (F1, F2) where the 64 modalities are plotted. This figure shows that the modalities vary according to a specific gradient going from the lowest to the highest modality values. As for example, the modalities of the SM3 * index follow a parabolic like curve from SM31 * to SM37 *; while the workload RAP * estimated by the pilot varies in the opposite direction (RAP1 * to RAP5 *).

The proximity of different modalities can be studied knowing that 2 modalities are as close to each other as their interrelation can allow.

The matrix of distances between variables, in the factorial space formed with the first 7 factors, enables only to look after the nearest neighbours of each workload rating evaluation. One can see that the higher pilot workload ratings have as nearest neighbours classes of indices corresponding to higher heart rate and lower heart rate variability and those of lower pilot workload ratings have only as nearest neighbours classes of indices corresponding to lower heart rates and higher heart rate variability.

In conclusion, we do think that heart rate and heart rate variability ambulatory monitoring of aircrews can be a good means to assess the impact and difficulty of task performance. From these last findings of a correspondence between heart rate variability and pilot ratings there is suggestive evidence of the possibility to include heart rate variability in a pilot workload model.

3. Dynamic Workload Modelling

The research results mentioned above suggested that workload ratings might be modelled using data extraneous to the pilot, such as aircraft and flight status measures. The research information reported in the previous paragraph illustrated however that heart rate data intraneous to the pilot may also be indicative of varying workload states. Hence the objective was formulated to attempt modeling ratings of the dynamic workload method by means of aircraft data, heart rate variability parameters and flight status measures. The study reported in this section was performed in cooperation with Dunlap & Associates and Cochlin Laboratory of Physiology. It utilizes data collected during 60 hours of actual route-proving flights in the A310 Certification campaign late 1982 and early 1983. The purpose of this study was to develop and validate a statistical model which would be capable of predicting the subjective mental workload ratings actually given by the pilots during the certification campaign. From over 60 hours of route proving flights on the European network, 31 flights averaging approximately one hour each were used to build a predictive model since aircraft data and heart rate measures on both pilots were available.

The task of developing a model capable of predicting a pilot's subjective workload rating involved both extensive data manipulation and management as well as the application of a rigorous statistical approach to avoid the possibility of deriving spurious results. Data management was a major undertaking both because of the size of the data sets and because four different sets of information recorded in completely different ways had to be integrated.

These were the pilot and observer ratings, the aircraft flight parameters and the pilot heart data. The fourth data set consisted of printed log sheets prepared during the various flights showing the flight phase, e.g. takeoff, climb etc., and flight condition, e.g. normal, emergency etc. Figure 14 shows the flow of these data through the various major processing steps. Development of the model was undertaken using a "split halves" design in which half of the data were used to construct the

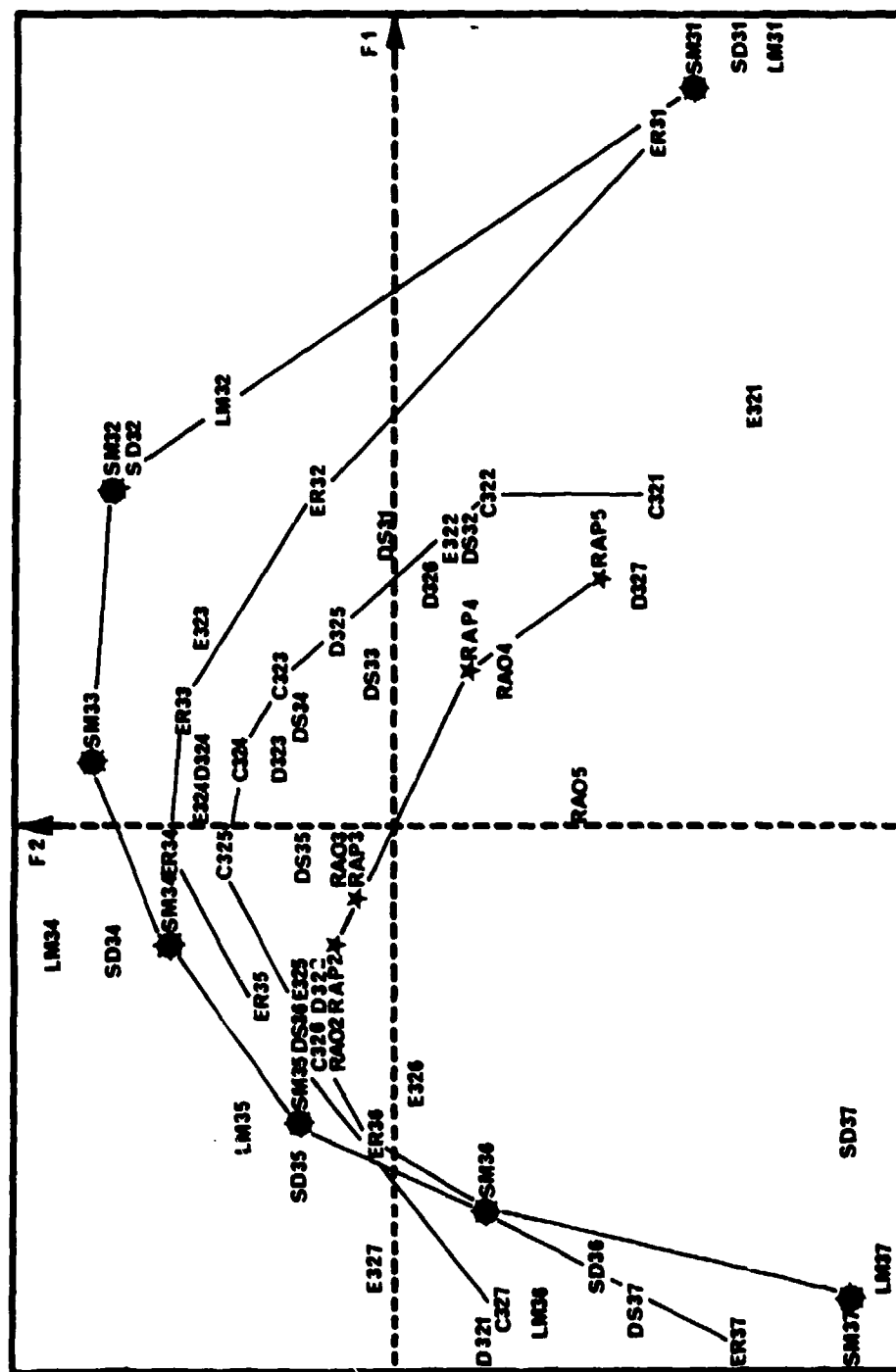


Fig. 13, Factorial space (F1, F2) of the 64 modalities
RAP modalities of pilot workload rating evaluations
SM modalities of the RMCSD heart rate index

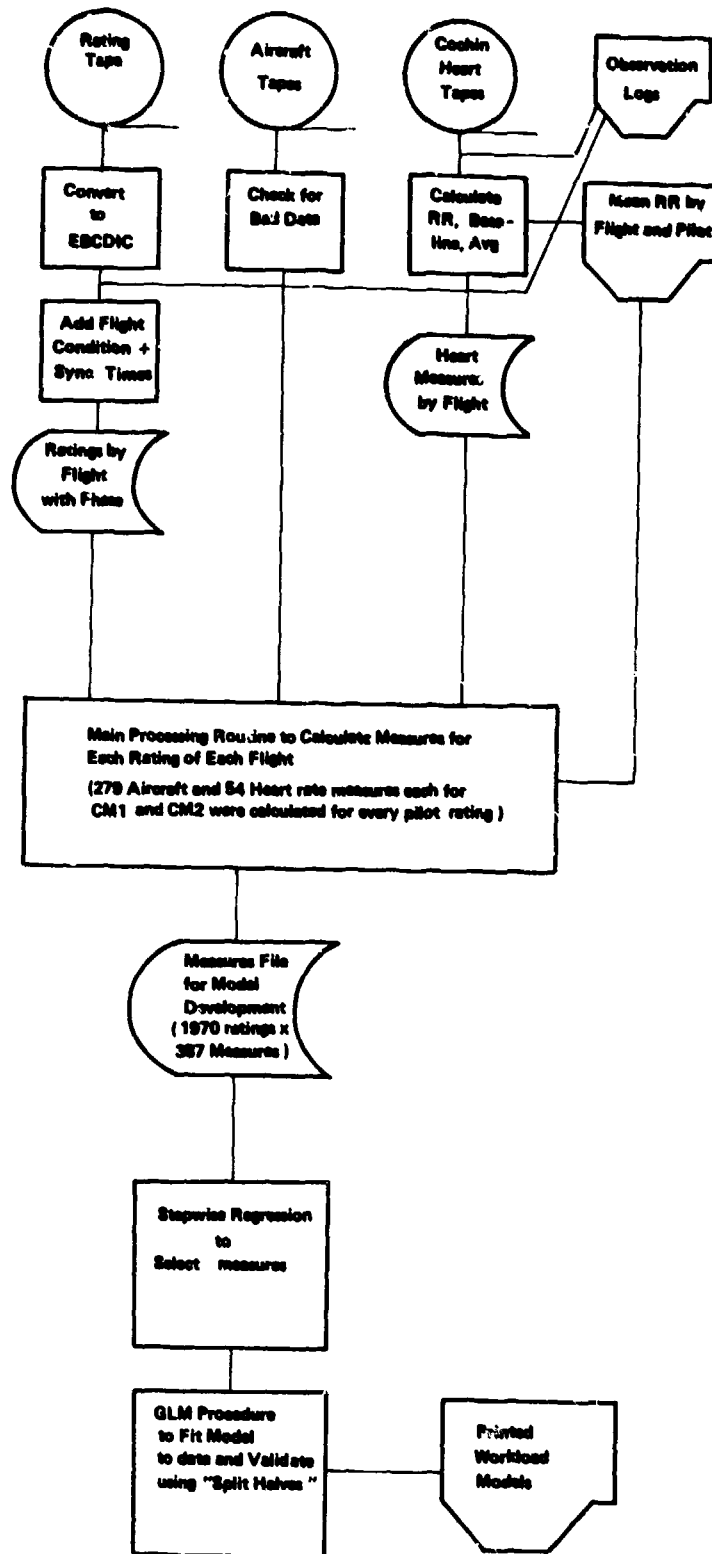


Fig.14 Data management flow

model and the remaining half were used as a reliability and validity test. The 7-point rating scale ranging from 2 through 8 was employed for the certification flights and used as the dependent measure in this study.

The model building half was used for the stepwise multiple regression screening of candidate measures previously mentioned.

Models were calculated using the General Linear Models (GLM) (3) (34), approach which permits the use of both continuous and discrete variables. Thus, measured data on heart rates and the aircraft (continuous measures) and categories of aircraft status information such as flight phase (takeoff, climb, approach, etc.) could all be used to predict the workload ratings which had been given by the 14 pilots during the flight.

The resulting model coefficients were utilized to calculate a predicted workload rating for each data point in the validity sample. The actual and predicted ratings were then correlated and the model was either accepted as valid or rejected based on the significance and magnitude of the correlation.

The application of these data management methods resulted in the calculation of over 50 different models of pilot workload.

The aircraft measures considered included:

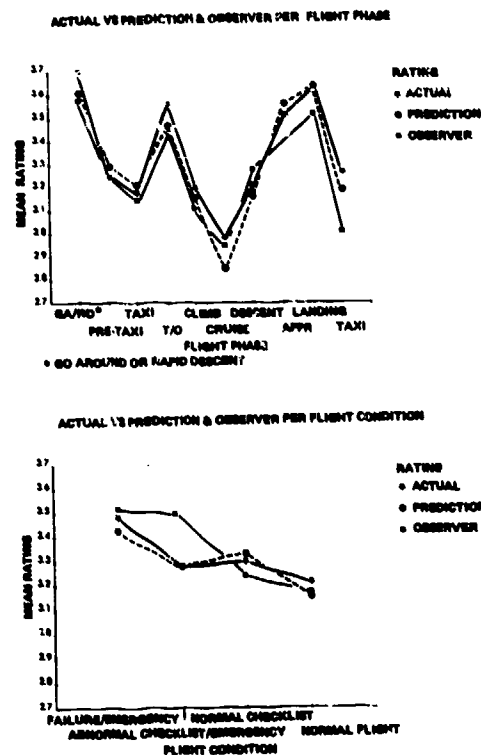
- exponential averages,
- rates through zero,
- reversal rates,
- number of AFC modes on.

The heart rate measures included:

- level
- difference (baseline, overall mean)
- trend (short, long terms)
- variance (short, long terms).

The resulting model finally selected as the best was proved reliable and a valid predictor of the rating pilots gave, significantly better than chance ($p < 0.0001$) (Multiple R of 0.67). Ratings predicted by the model correlated well with actual ratings, the model also predicting the ratings of the pilots more accurately than did the on-board observers.

This can be appreciated by flight phase and by scenario on figure 15. These graphs present mean ratings as a function of flight phase or flight condition and use:



- Aircraft measures
- Heart rate measures on rating pilot
- Heart rate measures on non-rating pilot
- Flight condition
- Flight phase
- Scenario
- Pilot flying/pilot non flying specification.

In general, this shows that higher mean workload ratings were associated with situations and conditions which research and experience suggested would show higher workload.

Regardless of the particular aspect of workload actually being addressed by the subjective ratings given by the A310 certification pilots, it was possible to utilize the data available to calculate a valid and reliable predictive model. Moreover, all three types of data (aircraft, heart rate and variability) play important roles in the model thereby further reinforcing the notion that workload is a "multi-dimensional mentally determined construct" (35).

PITFALLS AND LIMITATIONS

1. Dynamic Methods

Although pilot workload has been a concern for sometime, there has been little large-scale research conducted on commercial aircraft to date. The minimum crew certification programme for the AIRBUS INDUSTRIE A300 FF and A310 provided a unique opportunity to perform practical work in this field. Heart rate monitoring analysis could be done since measurement equipment was easy to install and unobtrusive when compared to other physiological measures. Subjective assessment of pilot workload was also easily accepted by those involved in the evaluation flights. Some pioneering work had already been done by Cooper and Harper (10) to develop a scale to rate aircraft handling qualities. The practice of opinion surveys well known in marketing and attitude research could readily be transferred to the cockpit area by adapting the scale just mentioned. However the apparent additional burden for pilots to have to rate their own workload prompted AIRBUS INDUSTRIE to adopt a cautious approach with regard to the amount of rating categories. At first, in the A300 FF, the pilot scale only contained 5 choices, which was eventually expanded to 7 for the A310 minimum crew campaigns, the former having proven not to be disturbing at all.

It does not go unnoticed that most ratings obtained with either scale had the propensity to be in the medium workload category. Rating distributions however clearly proved sensitive to workload alterations as shown earlier.

With the 7-point scale in the A310 campaigns we obtained even more continuously bell-shaped histograms whose median would systematically shift rightwards with increasing workload.

Finally, for the EFIS and FMS experiments a continuous 1 to 10 "bother" scale was also easily adopted by pilots. This scale again allowed for much more variation to be expressed.

It is also clear that the 5th (A300 FF) or 7th (A310) extreme workload category pushbutton on the rating box is intentionally missing. The reason being that in such a saturated situation the observer-rater would obviously not insist on a workload rating and rather call off the scenario being exercised to have pilots back to their primary concern i.e. maintaining flight safety. Because of this the A and E categories (A300 FF) or 2 and 8 ratings (A310) were to be considered as anchor points at both extremes of the workload range.

Having trained participating pilots to rate their workload evaluation with the scale almost continuously (at observer's request) we have to maintain that universal calibration is improbable be it only because everyone may be giving different attributes to the term workload, is having a different perception and is adopting a different attitude. Cumulating frequency or time distribution histograms may therefore appear as a simplistic accounting procedure whatever alternatives there may be. Not being engaged in pure scientifically-oriented research AIRBUS INDUSTRIE had purposely chosen not to impose any specific workload definition.

2. Workload Assessments

Most important was the ability to measure variations of reported workload and the possibility to express acceptability judgements throughout the scenario range. Workload in this sense is a human by-product resulting from a variety of man-machine and man-man information exchange processes based on the concepts of communication theory (36) (37). Boy expanded this kinetic interpretation of workload to several (workload) variability indices characterizing the informational entropy of crew-organizational perception and memorization procedures and decision strategies (38) (39). One should remind to stay clear from the temptation to make absolute quantifications of workload just as in thermodynamics it is not possible to measure entropy by means of a direct measuring equipment but such as a thermometer or a manometer.

Given the multitude of influences very precise quantifications of workload to assess the impact of a minor design change, the effect of a small procedural modification or the influence of an alteration in flight scenario may not necessarily make sense. No flight ever resembles any other as weather situation, ATC communication, air traffic, aircraft condition, crew contact and the operational context never are exactly the same without rigorous experimental precaution. It may be hazardous or even illusive to reproduce a flight for the sole purpose of absolute workload determination given the many influences that mediate the variation process and contribute to introduce biases and errors in such assessments.

3. Modelling Workload

What matters most, for an aircraft manufacturer is to be able to detect workload changes and trends as a function of evolving situations and resulting crew activity organisation within a given cockpit environment.

This is why experimental control precautions were taken in the performance criteria tests described earlier in Part I. Factorial designs for analysis of variance were rigorously adopted capable to counterbalance for confounding effects or to separate them from the effects of interest during analysis.

After the completion of most developmental studies, a central question concerns the practicality of the results. In most cases, research data are collected under conditions which permit an unusually high degree of control. The ability to extend results based on these data to "normal" conditions without the rigours of experimental controls is often an issue. However, in the context of the workload model derived from the A310 minimum crew campaign there is no such problem. The data for that study were collected during route-proving certification flights which were designed to be realistic. In fact the use of the data to support the construction of a model was not contemplated during the process of certifying the A310. It is therefore reasonable to conclude that the model developed in this study is realistic and representative of certification flights and, likely, the normal line operations these flights were intended to simulate.

The preliminary development study (discussed above) also found a strong predictive model of pilot ratings made on a 10 point scale not unlike the 7 point version used during the A310 flights. The preliminary study, however, was conducted in an A300 test aircraft. This provides the suggestion that models can be developed which would be valid across a wide variety of aircraft types. Indeed, there is nothing inherent in any of the measures used in the model which would suggest that they were not widely applicable to jet aircraft with similar performance characteristics. The GLM modelling technique would allow the aircraft type to be used as a classification variable if similar data across the aircraft types were available.

The model (discussed above) involved numerous scenarios which covered a great variety of normal, abnormal and emergency operating conditions. This model seemed quite capable of tracking the subjective judgements of the pilots across this range of circumstances. Thus, the flying task need not be kept uniform in order to be able to predict pilot ratings.

The proved validity, reliability and realism of the model does not, necessarily, insure its utility to AIRBUS INDUSTRIE. The model was developed using the A310 (200 series) and flights with a duration of approximately one hour. Its universality has yet to be demonstrated and validation work with other flight measurements on other AIRBUS-versions is underway. Nevertheless, there is ample evidence that the approach employed and, perhaps, the basics of the model could have widespread applicability.

CONCLUSION

Dynamic assessment of workload had never been performed by a European aircraft manufacturer to the extent it is reported in this paper.

Clearly, the incentive was to certify the advanced cockpits of new technology aircraft with a crew complement of two pilots. But beyond that it brought a realization that meaningful human factors work can make sense if proper measurement procedures are used and if now available computational facilities are utilized. In particular, while not overstressing the merits of subjective (workload) rating it helped once again to accept that human judgement can be relied on if and when used with precaution. Moreover, it was extremely encouraging to see the work on ambulatory monitoring of heart rate to come to fruition also partly because many attempts in the past led to ever-increasing scepticisms with regard to this field.


Regardless of the particular aspect of workload actually being addressed by the subjective ratings given by the A310 certification pilots, it was possible to utilize the available data to calculate a valid and reliable predictive model. The existence of the model is, by itself, a significant finding. The complete analysis of all the dimensions of the model and its potential implications for the theory of workload and its measurement were well beyond the realm of the present paper.

It is worthy of note, however, that the process of developing this model has shown that there is an underlying order to the dynamic workload assessments performed by pilots in situations such as Certification flights. The ability to detect workload variations and to predict the subjective ratings opens up numerous possibilities for additional research and development in other areas than the certification of aircraft.

REFERENCES

- 1 BILLINGS C E
CHEANEY E S Information transfer problems in the aviation system. NASA Technical Paper 1875, 1981.
- 2 JAHNS D W Operator workload. What it is and how it should be measured? Crew System Design. Proceedings of an Interagency Conference on Management and Technology in the crew systems design process, Los Angeles, California, September 1972.
- 3 GARTNER W B
MURPHY M R Pilot Workload and Fatigue: A critical survey of concepts and assessment techniques NASA TN D-8365, 1976.
- 4 SHERIDAN T B
SIMPSON R W Towards the Definition and Measurement of the Mental Workload of Transport Pilots. Final Report, Massachusetts Institute of Technology Contract DOT -OS -70055, January 1979.
- 5 KATZ J G Pilot Workload in the Air Transport Environment : Measurement, Theory and the Influence of Air Traffic Control. FTL Report R 80 - 3, M.I.T., May 1980.
- 6 WIERVILLE W W Physiological measures of aircrew mental workload. Human Factors 21 (5), 575-593, 1979.
- 7 ROSCOE A H (Ed.) Assessing pilot workload. AGARDograph No. 233, February 1978.

- 8 LAUBER J K Resource Management on the Flight Deck. Proceedings of a NASA/Industry Workshop held at San Francisco, California. Report NASA CP-2120, March 1980.
- 9 BERLINER C
ANGELL D
SPEYER J W Behaviours, measures and instruments for performance evaluation in simulated environments. Symposium and workshop on the Quantification of human performance. M-3, 7 Subcommittee, Electronic Industries Association, 1964.
- 10 CONNER G E
HARPER R P Jr The use of pilot rating in the evaluation of aircraft handling qualities. Report NASA TN D-5153, 1969.
- 11 McCORMICK E J Human Factors in Engineering and Design. McGraw-Hill, Inc. New York, 1976.
- 12 DUBOIS Prof J B Direction des Recherches. Recueil de Donnees et Moyens d'Essais, Ergonomiques Tome 1 et 2, Doc. A.A. 13/74, 1974.
- 13 SPEYER J J
FORT A P Static Taskload Analysis DC-9/A300 FF Comparison Certification Report AIRBUS INDUSTRIE AI/V-F 1304/81, 1981.
- 14 SPEYER J J
FORT A P Static Taskload Analysis A300 FF/A 310 Comparison Certification Report AIRBUS INDUSTRIE AI/V-F 020/83, 1983.
- 15 LEBART L
FENELON J P Statistique et Informatique Appliquees, 3eme edition, DUNOD, 1975.
- 16 SEIGEL S Non Parametric Statistic for the Behavioral Sciences. McGraw-Hill, New York, 1956.
- 17 SPEYER J J
FORT A P Performance Criteria Analysis: Evaluation of EFIS and FMS Certification Flight Test Report. AIRBUS INDUSTRIE, AI/V-F 026/83, 1983.
- 18 NIE N H
et al Statistical Package for the Social Sciences, 2nd edition. McGraw-Hill, New York, 1975.
- 19 SPEYER J J
FORT A P Dynamic Workload Analysis Simulator Campaign A300 FF Certification Flight Test Report. AIRBUS INDUSTRIE, AI/V-F 1305/81, 1981.
- 20 SPEYER J J
FORT A P Dynamic Workload Analysis Flight Campaign A300 FF Certification Flight Test Report. AIRBUS INDUSTRIE, AI/V-F 1306/81, 1981.
- 21 SPEYER J J
FORT A P Dynamic Workload Analysis Simulator Campaign A310 Certification Flight Test Report. AIRBUS INDUSTRIE, AI/V-F 022/83, 1983.
- 22 SPEYER J J
FORT A P Dynamic Workload Analysis Flight Campaign A310 Certification Flight Test Report. AIRBUS INDUSTRIE, AI/V-F 024/83, 1983.
- 23 HART S G
CHILDRESS M E
HAUDER J R Individual Definitions of the Term "Workload". Paper presented at the 1982 Psychology in the DOD Symposium.
- 24 SPEYER J J
FORT A P Certification experience with methods for minimum crew demonstration. AGARD Conference Proceedings No. 347 AGARD Paris 1983.
- 25 SPEYER J J
FORT A P Performance Criteria Analysis: Evaluation of EFIS and FMS Certification Flight Test Report. AIRBUS INDUSTRIE, AI/V-F 026/83, 1983.
- 26 BLOMBERG R D
PEPLER R D
SPEYER J J Performance Evaluation of Electronic Flight Instruments Second Symposium on Aviation Psychology, Columbus, Ohio 26-27 April 1983.
- 27 TEKAIA F
et al Incidence des contraintes Psychiques et Intellectuelles sur la Frequence Cardiaque, les Cahiers de l'Analyse des donnees, Vol VII - 1981 no 2 p. 175-185.
- 28 FOUILLOT J P
REGNARD J
SPEYER J J
FORT A P Heart Rate Monitoring Analysis A310 Flight Campaign. A310 Certification Flight Test Report. AIRBUS INDUSTRIE, AI/V-F 304/83, 1983.
- 29 FOUILLOT J P
et al Ambulatory Monitoring of Air Crew Heart Rate Variability. Paper presented at the 1985 ISAM Symposium. Proceedings of the Fifth International Symposium on Ambulatory Monitoring.
- 30 TEKAIA F
et al The Nearest Neighbours : Application to Workload and Heart Rate Variability Op. Cit.
- 31 BENZECRI J P Pratique de l'Analyse des Donnees, Analyse des Correspondances, DUNOD 1980.
- 32 FOUILLOT J P
et al Methodology of Heart Rate Ambulatory Monitoring Recordings Analysis, in relation to activity: Applications to Sports training and workload studies. ISAM - GENT, Belgium 1981 pp. 377-383. Proceedings of Fourth International Symposium on Ambulatory Monitoring, F D Scott, Editor, Academic Press 1982.
- 33 NETER J
WASSERMAN W Applied Linear Statistical Models, Homewood, Illinois. Irwin 1974.

- 34 ANON SAS Institute. SAS Users Guide : Statistics. 1982 Edition SAS Institute cary North Carolina.
- 35 O'DONNELL R Conceptual Framework for the Development of Workload Metrics in Sustained Operations. AFAMRI, Wright-Patterson AFB, 1983.
- 36 WIENER N Cybernetics . or Control & Communication in the Animal and the Machine. MIT -Press, Cambridge, Massachussetta, 1948.
- 37 WEAVER W SHANNON C E The mathematical theory of communication. University of Illinois, 1949.
- 38 BOYGA MESSAGE : An Expert System for Aircraft Crew Workload Assessment. CERT-ONERA, Toulouse, France
- 39 BOYGA Le systeme MESSAGE : un premier pas vers l'analyse assistee par ordinateur des interactions homme-machine. Le Travail Humain, tome 46 no 2, 1983
- 

CHAPTER 15

MEASUREMENT OF PILOT WORKLOAD

by

Sandra G Hart
NASA-Ames Research Center
Moffett Field
California 94035, USA

INTRODUCTION

Pilot workload may be defined as the cost incurred by the human operators of complex airborne systems in accomplishing the operational requirements imposed on them. If all pilots could perform all flight-related activities on time and without error, and if they could do so using available hardware, software, and human resources, the concept of workload would have little practical significance. However, they often cannot. Automation has been offered as a solution to an increasing number of workload-related problems in existing systems or predicted for those under development. In addition, there has been an ever-increasing tendency to reduce the number of crewmembers in aircraft cockpits. Again, automatic subsystems are provided to moderate the demands thus placed on the remaining crewmembers. Attempts to completely replace humans by automatic systems have failed, however, because human capabilities, adaptability, and flexibility continue to surpass those of the most advanced and sophisticated systems.

To achieve the desired levels of overall system effectiveness, aircraft must be designed that take advantage of the capabilities of the remaining crewmembers and impose acceptable levels of workload. Thus, the concept of workload has received an increasing amount of theoretical attention during the past decade and it has become an important consideration in system design. This interest has been prompted by the realization that the human element in advanced man-machine systems represents the limiting factor in accomplishing, increasingly complex activities.

In some cases, apparently human limitations reflect the consequences of controls, displays, and automatic subsystems that are poorly designed or that are poorly interfaced with the pilots. In other cases, the demands imposed on the pilots exceed their capabilities either momentarily or for extended periods. Finally, the environments in which some tasks are performed impose additional demands on the pilots, combining with other sources of workload to exceed their capabilities.

SOURCES OF WORKLOAD

The relationship between workload, human behaviour and system performance is complex. Thus, measurement procedures that are inappropriate, insensitive, or simplistic may provide trivial or misleading answers. The components of workload for different activities vary and the workload experienced by individuals faced with apparently identical task requirements may be quite different. To some extent, this occurs because the workload of a task is not uniquely defined by its objective demands; it also reflects an operator's responses to them as well. In addition, various measures may provide different workload estimates for the same task because they reflect unique aspects of it, the circumstances in which it is performed, and individual differences in behavior and experience. Thus, the utility of the information that measures provide may vary with the situation under consideration. The factors that contribute to pilot workload include the demands imposed by the task, the available system resources, the environments in which it is performed, and individual differences among pilots.

IMPOSED DEMANDS

The demands that are imposed on pilots are created by *what* they are asked to achieve (eg the objective goals of the flight and requirements for speed and precision) and *when* (eg schedules, procedures and deadlines). Some flight tasks are intrinsically more demanding than others, and the difficulty of almost any task can be altered by a requirement for additional speed or accuracy. The system resources that are provided define *how* the pilots can accomplish the task demands.

They include controls, displays, automatic sub-systems, other crew members, and ground support. Poor display design, inaccessible controls, poor handling qualities, and too much or too little information can increase workload, even for flight tasks that might otherwise impose relatively low demands. Finally, *where* a task is performed (eg geographical location, altitude, time of day, weather) may also affect workload. For example, visual workload may be increased by poor visibility, physical workload may be increased by turbulence, and threats from natural or man-made sources certainly increases stress-related components of workload. These elements may act independently to create the workload level that is imposed on a pilot or they may interact, enhancing or mitigating each others' effects.

EXPERIENCED DEMANDS

Finally, *who* performs the task determines the actual level of workload experienced by a particular pilot. Most tasks require certain basic skills, knowledge, and training; unskilled or inexperienced pilots experience higher levels of workload than more skilled or experienced pilots. In addition, incorrect strategies, insufficient effort, or pilot errors may result in higher levels of workload associated with detecting, resolving and recovering from the problems created by the pilots themselves. Finally pilots' expectations, previous experiences, and physical and emotional states can affect their subjective experiences and evaluations of workload; as well as their performance. Thus, the "work" that is "loaded" on a pilot is an important component of the workload experienced by a particular pilot, but the demands experienced during a specific flight may reflect a number of other factors as well.

AD-P005 642

Different types of questions might be asked about workload. The underlying motives might be economic, political, engineering design, safety, or humanitarian. The goal might be to prevent potential problems or to identify and solve those that already exist. Some questions relate to the sources of workload (eg the effects of specific flight tasks, procedures, schedules, alternative types of controls and displays, the addition of automated sub-systems, degraded flight modes, and pilot selection and training). Others focus on the consequences of inappropriate levels of workload (eg the likelihood of fatigue, performance decrements, or health problems). Yet others relate to the relative merits of alternative solutions to workload problems (eg modified system designs, mission requirements, or crew complements). Finally many questions are posed about the extent to which a pilot's "resources" are consumed by the nominal flying task and are, therefore, unavailable for additional tasks or emergency situations. Different questions demand different procedures to provide a valid and practical measure or solution.

WORKLOAD MEASURES

Despite its complexity, workload is assumed to be an important and practically relevant entity and a number of valid, sensitive, and reliable measurement techniques have been developed. Workload measures are usually organized into four categories: (a) objective measures of primary or secondary task performance, (b) subjective ratings, (c) physiological recordings, and (d) analytic techniques. Each type of measure has advantages and disadvantages and limitations in range of activities and questions to which it applies; the evidence they provide may or may not be useful, depending on the situation.

A structure and rationale for selecting and applying workload measures and interpreting the results relies on a theoretical understanding of the potential contributors to pilot workload and a precise definition of the goal of a specific analysis. For example, questions about task demands might be addressed by analytic procedures (eg task and time-line analysis). Questions about control and display design might be addressed by behavioral measures (eg reaction time, accuracy, eye point of regard), physiological measures (evoked cortical potentials), pilot opinion, and models of human operator control, attention, and decision making. Questions about the effect of the environment on workload might be addressed by measures of physiological arousal (eg heart rate, respiration) and pilot opinion. Finally, questions about reserve capacity are often answered with secondary-task techniques.

It is difficult to measure workload absolutely. To some extent, this occurs because the workload of different tasks is created by different factors. Thus, the values obtained with the same measure used in different situations may reflect different phenomena. A workload rating for one task might reflect the level of time pressure experienced, whereas another, apparently similar rating, might represent mental effort or stress. An increase in heart rate might reflect the stress of low level flight or the physical effort required to control an aircraft in heavy turbulence. Each evaluation reflects the cost incurred in performing the task, but the information provided by the measures is not equivalent. Furthermore, it may be difficult to compare workload estimates obtained with different measures directly.

For this reason, most workload evaluations are relative; one flight segment is compared to another, a new aircraft is compared to a reference aircraft, alternative display designs are compared to each other, a degraded environment is compared to the nominal case, or the workload of a skilled pilot is compared to that of a novice. In each case, it is assumed that the salient features of the activities are roughly equivalent, except those that are experimentally manipulated. Thus, other, irrelevant, variables are held constant, information obtained about the variables of interest can be compared directly, and the reference task or configuration provides a context within which the results can be interpreted.

EXAMPLE OF A WORKLOAD ANALYSIS

A standard task has been provided by the AGARD FMP Panel for which a candidate workload assessment procedure is to be recommended: the final five minutes prior to landing for a jet transport (Appendix 1). The task requirements include manual altitude, speed and flight path control, navigation (using the Instrument Landing System — ILS), communications, checklists, instrument checks, and callouts. The approach is flown in the rain with a 200ft cloud base and limited visibility in a typical aircraft configured for two pilots. No system failures are encountered nor are any modifications made to the intended flight plan.

DEFINING THE QUESTION

Different questions might be asked about the workload of this flight segment:

(a) Is the allocation of duties between the two crewmembers optimal? (b) What is the effect of degraded weather during approach and landing? (c) Could pilot workload be reduced by automatic altitude callouts or checklists? or (d) Are there momentary workload levels that are too high? I will focus on the first question for this paper study. To answer this question, information is needed about the tasks each pilot is expected to do, when he must do them, and the relative amounts and types of workload the pilots will encounter during different approaches. For example, the pilot-flying might experience continuous visual and manual workload while the pilot-not-flying might experience high levels of monitoring and communications workload. Furthermore, differences in responsibility between the right and left seats might create relatively subtle differences in workload from the pilot's perspectives.

GENERAL PROCEDURE

In this section, I will describe how the workload analysis will be structured and an inflight experiment conducted. First, the activities required of the crew as a team must be defined and a nominal time-line for these activities established. Next, the distributions of duties adopted by individual crews (and the resulting workload levels) must be assessed inflight. The former is accomplished analytically, the latter, empirically. The preliminary analysis provides a structure for the subsequent inflight

experiment. It suggests how to segment the flight for analysis, when to apply workload measures, and candidate tasks for a detailed analysis.

The pilots will perform the approach and landing in a standard cockpit with all equipment functioning normally. Since no alternative controls, displays, levels of automation, or crew sizes will be considered, the effect of system resources upon workload will not be addressed. Likewise, since the approaches will be flown under identical meteorological conditions, the influence of environment on workload will not be examined directly.

Criterion performance levels are established for airspeed (± 3 kts), rate of descent on the glideslope (± 50 ft/min), localizer tracking (± 2.5 deg) and touchdown point (within 100m and at less than 6 ft/sec). In addition, a list of discrete activities that must be accomplished during each segment (eg callouts, flap settings, landing checks, communications), will be prepared.

Qualified transport pilots will participate in the flight experiment, using equipment and procedures with which they are familiar. Flight time and familiarity with the aircraft and routes will not be experimentally manipulated.

SELECTION OF MEASURES

Task Analysis/Time Line

A task analysis will provide information about what activities are required while a time-line will establish the schedules, procedures, and deadlines. Some preliminary information about the workload imposed by each task and the time it requires can be obtained from existing data bases (eg Hart & Bartolucci (1)).

The entire five-minute flight could be evaluated as a single entity, however, subdividing it into four intervals allows a more precise and diagnostic assessment. The activities performed during each segment (Table 1) include: flight-path control, navigation, communications, checklists, crosschecks, or callouts, and discrete actions. The segments represent meaningful units of activity from a pilot's perspective rather than equal intervals of time.

Measures of Performance

Primary Task Compliance with target performance values will be evaluated at 30-sec intervals by a cockpit observer. He will also record when discrete actions are performed and by whom. Two measures of performance will be obtained that are often sensitive to inflight workload: flight path control (glideslope and localizer deviation) and communications. Control measures provide an objective summary of how well the pilots manage an aircraft to achieve a smooth and precise approach. Deviations during any 30-sec interval will indicate periods of time when the pilot-flying was sufficiently overloaded by other actions that primary flightpath control suffered. A communications analysis will provide an objective estimate of ATC-related workload levels. This is possible because standardized taxonomies of communications exist in which a priori estimates of the workload imposed by communications tasks have been quantified (1), (2), (3)

TABLE 1. Segments of flight for workload analysis.

Segment 1:	Descent from 4000 ft to level off at 2000 ft
	a. Reduce speed from 250 kts to 210 kts using speedbrakes
	b. Approach checklist
	c. Radar vectors to intersect ILS
Segment 2:	Level at 2000 ft to glideslope capture
	a. Reduce speed to 140 kts
	b. Gear down
	c. Flaps to 1, 5, then 15
	d. Set altimeters
	e. Localizer intercept
	f. Change to tower frequency
	g. Landing checklist complete
Segment 3:	Glideslope capture to touchdown
	a. Reduce speed to VAT+10
	b. Descend on glideslope
	c. Landing flaps selected
	d. Final landing information obtained and checked
	e. Altitude callouts
Segment 4:	Touchdown to taxi off runway
	a. Reverse thrust, deceleration
	b. Braking
	c. Nose wheel steering
	d. Change to ground frequency

Secondary Task Most secondary task measures of pilot workload are inappropriate inflight because they are difficult to implement and might compromise safety. However, interval production is one exception because stimuli can be presented and responses obtained with minimal instrumentation and it does not intrude on primary task performance. When workload levels become very high, it is performance on the interval production task that suffers rather than aircraft control. In fact, pilots may simply forget they are in the midst of producing an interval when overload situations occur. The occurrence of such "timeouts" can be evaluated as indicators of workload peaks. Furthermore, previous research has shown that this measure is sensitive to the workload levels encountered in different segments of simulated flight (4), (5), (6), (7). Because it is difficult to concentrate on the passage of time in the presence of any other activity, clock time continues but subjective awareness of it may not, leading to an underestimation of the passage of time (e.g. longer production intervals and shorter verbal estimates) and increased variability.

In this flight, as in many previous simulations, 10-sec intervals will be selected for the interval production task. At previously established points in each of the four segments of flight, the observer will ask the pilots to start a timer mounted on the outboard side of their seats, wait until they feel that 10 sec has elapsed, and then stop the timer. The observer will collect the timers, record the produced durations and replace the timers for the next interval production. In order to avoid interfering with flying at critical times, interval productions can not be recorded throughout the flight.

However, the information that they provide when they are given indicates the relative amounts of mental workload experienced by each pilot at that instant. To control for individual differences in timing, baseline productions will be obtained prior to the flight and measures obtained inflight will be expressed as deviations from these values.

Physiological

Two physiological measures will be obtained for each analysis segment: heart rate and heart rate variability. These measures reflect several factors that can contribute to flight-task workload: stress, responsibility, physical effort and mental effort. Physiological measures generally have the advantage of being unobtrusive. That is, they can be obtained without requiring attention from the pilot or interfering with the flight. In addition, since they can be recorded relatively continuously, they can reflect momentary fluctuations in workload. Finally, they provide an objective indication of involuntary physiological changes that often accompany variations in workload. The disadvantages include a lack of diagnosticity. That is, most physiological measures reflect non-specific responses to many sources of stress. These responses may be due to the demands imposed by the flight, the environment, or the pilot, or to other factors that are less directly related to workload. Cardiovascular responses do, however, provide an integrated indication of the total impact of the flight on the pilots that does not also reflect the characteristics of the system (as many performance measure do) or the pilots' biases and misconceptions (as subjective ratings do).

As the heart muscle tenses and relaxes, circulating blood through the system, variations in the sound of the heart beat and residual electrical potential can be recorded on the skin. These electrical signals can be recorded with a portable biomedical monitoring device such as the Vitalog. The Vitalog is the size of a pocket calculator and can be worn in the pocket of a pilot's flight suit (8). Three electrodes are attached to the pilot's chest with electrode paste and adhesive tape. The Vitalog detects "R-waves" and records the average inter-beat interval and with a very high sampling rate (20 times per second), also provides information for the proposed analysis of heart rate variability.

Heart Rate

The average beat-to-beat interval has been shown to reflect the stress associated with specific flight-related activities. In general, the expectation is that heart rate will increase as workload is increased. For example, Hart, Hanser and Lester (8) and Roscoe (9) found that heart rates are typically elevated during take-off and landing and return to baseline levels at altitude. In addition, substantially greater increases were found for the pilot-flying during take-off and landing than for the pilot-not-flying. It is possible that the feeling of responsibility and level of preparedness that must be maintained by the pilot-flying could result in their elevated levels of arousal. Thus, heart rate measures should be able to differentiate between two crew members.

Heart rate may not be sensitive to differences in mental workload, however.

For example, it has been found to be relatively insensitive to the workload of tasks performed in a laboratory or simulator when the sources of workload were primarily mental and the stress associated with flight was not present (10), (11). Thus, heart rate provides information about pilots' arousal levels, but may not relate to other aspects of workload.

Heart Rate Variability

A second cardiovascular measure will be used that has been found to reflect even subtle variations in mental workload; heart rate variability or sinus arrhythmia. The general finding has been that heart rate irregularity decreases as the difficulty of a task is increased. The specific technique proposed is based on an idea suggested by Mulder (12) that controlled or attentive cognitive processing may lead to a "defense reaction" that is initiated by an increase in effort and reflected in a decrease in heart rate variability. This is manifested in a reduction in the amplitude of the 0.1 Hz component of the frequency spectrum of beat-to-beat intervals. Mulder's analytic technique was based on aggregates of 256 heart rate samples, however, (about 4 min for a heart rate of 65 beats/min), which would not be precise enough for current application (several segments will last less than 1 min).

Moray and his colleagues at the University of Toronto have developed an alternative method of obtaining an estimate of the power in the 0.1 Hz region of the frequency spectrum that looks very promising (13), (14). They developed a "black box" that monitors, records, and quantifies this cardiovascular measure virtually continuously, providing a sensitive real-time indication of workload variations associated with difficulty manipulations within tasks and of the workload reduction that accompanies training. Inflight the information for this analysis could be recorded and stored for later, offline, analysis.

Subjective Ratings

Subjective ratings may come closest to tapping the essence of workload and provide the most generally applicable and sensitive measure. This is because they provide a direct indication of the impact of flight-related activities on pilots that integrates the effects of many workload contributors. Another advantage is that the pilots can let their experiences influence their judgements, thereby taking into account whatever they considered relevant in a particular flight segment. The disadvantage is the potential for high levels of between-rater variability. Since the requirement to quantify one's experiences with experimentally-imposed rating scales is not a natural activity, there may be discrepancies between pilots' subjective experiences and their abilities to express these experiences with a specific rating scale. However, well-designed rating scales with operationally defined terms can resolve many potential problems.

Despite inconsistencies in the absolute values given with rating scales, the typical finding is that the rank-ordering of tasks or flight segments with respect to workload is quite consistent across raters. However, because the factors that contribute to workload vary between tasks and between raters, a multi-dimensional approach may be better able to capture all potentially relevant factors. The typical finding is that people can estimate specific components more accurately and consistently than they can the more global construct of workload and that they can evaluate experimentally relevant factors even though they might not have considered them in a global workload rating. The subscales must include questions about the objective demands imposed on pilots as well as their behavioral and emotional responses to them, but they must not be so numerous that they cannot be obtained inflight with minimal interference.

A rating scale has been developed at NASA-Ames Research Center that provides an overall workload score based on a weighted average of magnitude ratings on six subscales: Mental Demands, Physical Demands, Temporal Demands, Own Performance, Effort, and Frustration. The subscales were selected after a multi-year research effort, summarized in Hart and Staveland (15). The importance of each factor as a source of workload for a particular task is obtained by a simple pair-wise comparison among the six factors. Ratings on each subscale are obtained after each performance of the task. By giving more weight to ratings of factors that were most important during a particular task, the sensitivity of the derived workload score is, thereby, enhanced. The derived workload scores have substantially less between-rater variability than unidimensional workload ratings and the subscales provide diagnostic information about the specific sources of loading.

The first dimension of this two-dimensional rating scale (Importance) reflects the contribution of each factor to the workload of a specific task from the perspective of the pilot. This dimension is reflected in the weight given to each factor by the raters. The weights account for two potential sources of rating variability: differences in workload definitions between raters within a task and differences in the sources of workload between tasks. In addition, the weights also provide diagnostic information about the nature of the workload imposed by different tasks or experienced by different pilots. There are 15 possible pairwise combinations of the 6 scales. The number of times each factor is selected as being more relevant to the workload of a particular task, in comparison to each other factor, is tallied. The minimum tally for each factor is 0 (not at all relevant). The maximum tally is 5 (more important than any other factor).

The second dimension (Magnitude) reflects the numerical values given to each factor during or following performance of a task or task segment. Ratings are obtained for each scale individually. The scales are presented on a computer display or rating sheet. Responses are made with an analog input device, marking on the rating sheet, or verbally. Inflight, rating sheets or verbal responses are most practical. Each scale is presented as a 12-cm line divided in 20 equal intervals anchored by bipolar descriptors appropriate for that factor (e.g., Extremely Low/Extremely High). The responses are quantified on a scale from 1-100 in increments of five points during data analysis. The weights and ratings may or may not covary. For example, it is possible for mental demands to be the primary source of loading for a task, but the magnitude of those demands might be low. Conversely, the time pressure under which a task is performed might be the primary source of workload and the time demands might be rated as high.

The overall workload score is computed by multiplying each rating by the weight given to that factor by each subject. The sum of the weighted ratings for each task or task segment is divided by 15 (the sum of the weights). Table 2 depicts the procedure for computing a derived workload score. Sample weights and ratings are listed for an approach segment flown on autopilot with high planning and information-seeking demand and moderate time pressure and stress.

TABLE 2: Hypothetical example of weights and ratings given by a pilot during an approach and the derived workload.

Factor	Weight	Rating	Product (W*R)
Mental Demands	5	65	325
Physical Demands	0	10	0
Temporal Demands	3	60	180
Own Performance	1	50	50
Effort	3	45	135
Frustration	3	30	90
Sum			780
Derived Workload Score			52

Since the ratings are given very quickly, it is possible to obtain them inflight. The observer could hand pilots a rating sheet at the end of each segment. However, since the last two segments are relatively brief and giving ratings might interfere with safety of flight, these segments will be rated immediately after the aircraft arrives at the gate. Previous simulation and inflight research has shown that little information is lost when ratings for certain segments are not given immediately (8). An alternative method would be to obtain subjective ratings for all four segments during a post-flight debriefing. For this technique to be effective, however, a video-taped replay of the pilot's activities during each segment of flight should be provided as a mnemonic aide that can be stopped after each segment to obtain ratings. A high correlation has been found between "online" ratings and those obtained retrospectively with a visual recreation of the task (10), (16).

SUMMARY

A multi-stage process for evaluating the workload of a five-minute segment of flight including approach and landing for a typical transport aircraft was described. The goal of the analysis was to compare the workload of the two pilots. Four types of measurement techniques were suggested: 1 Analytic (a preliminary task and time-line analysis identified task requirements and target performance levels); 2 Performance (flight-path control, communications, and interval production); 3 Physiological (heart rate and heart rate variability); and 4 Subjective ratings (a multi-dimensional technique developed at NASA-Ames Research Center).

Different information about the research question is provided by each stage of the analysis procedure. The task and time-line analysis provides explicit information about what is expected of the pilots and when each subtask is to be performed. It can provide a priori estimates of workload and an organizational structure within which the information provided by the other measures can be related and interpreted.

The flight-path control measure of performance reflects the degree to which the pilot-flying was able to accomplish the primary control task. The types of communications tasks performed by the pilot-not-flying provide an independent estimate of his workload. In addition, errors and delays in response might indicate the presence of high workload levels. The accuracy and variability of time productions indicates the relative levels of mental workload experienced by the two pilots by reflecting the amount of available attention each was able to focus on the timing task.

Heart rate reflects the different levels of arousal experienced by the two pilots during each segment. Heart rate variability reflects the moment-to-moment cognitive demands placed on them within each segment.

The importance placed on each of the rating subscales reflects the differential sources of loading placed on each pilot. The numerical ratings reflect the magnitudes of the different types of loading for each pilot and flight segment. The derived workload score provided an integrated estimate of the overall workload imposed on each pilot, taking into account the fact that they are likely to encounter different sources of loading within and between segments.

By analyzing all of the information obtained inflight, and by comparing it to a priori estimates, a fairly complete picture of the sources and magnitudes of workload imposed on each pilot can be obtained and compared. This information might be used to identify moments in which one or the other pilot was over- or under-loaded and suggest a redistribution of duties or modified procedures. Given the information available about the specific nature of the tasks performed by each pilot, and the workload associated with each one, the decision of what modifications might be made could be accomplished with some assurance and the outcome of the modification could be predicted in advance.

Workload is a complex, multi-dimensional experience that reflects the cost to humans of accomplishing different tasks.

Although its definition might vary from one activity to the next, or from one person to another, it is a practically relevant and measurable quantity. By understanding the levels and types of workload imposed on pilots by different airborne systems and tasks, the quality of system design can be improved for aircraft under development and many operational problems can be resolved in existing aircraft. This only can be accomplished by selecting valid and reliable measures that address the type of question that has been posed.

REFERENCES

- 1 HART S G Pilot errors as a source of workload. *Human Factors* 25, 545-556, 1984
BORTOLUSSI M
- 2 SINGLEDECKER C A Subsidiary radio communications tasks for workload assessment in R & D simulations: I.
et al Conceptual Development and Task Workload Scaling (AFAMRL-TR-80-126) Wright
 Patterson AFB, OH, 1980
- 3 SILVERSTEIN L D A comparison of analytic and subjective techniques for estimating communications-related
GOMER F E workload during commercial transport operations. (NASA CR-2341) Washington, D C:
CRABTREE M S National Aeronautics and Space Administration, 1984
ACTON W H
- 4 BORTOLUSSI M R Measuring pilot workload in a motion base trainer: A comparison of four techniques.
KANTOWITZ B H Proceedings of the Third Biannual Symposium on Aviation Psychology. Columbus, OH:
HART S G Ohio State University, 263-270, 1985
- 5 GUNNING D Time estimation as a technique to measure workload. Proceedings of the Human Factors
 Society 22nd Annual Meeting, Santa Monica, CA: Human Factors Society, 41-45, 1978

- 6 HART S G
MCPHERSON D
LOOMIS L L Time estimation as a secondary task to measure workload: Summary of research. Proceedings of the 14th Annual Conference on Manual Control. Los Angeles, CA: University of Southern California, 693-712. 1978
- 7 WIERWILLE W W
CONNOR S A Evaluation of twenty workload assessment measures using a psychomotor task in a motion-base simulator. Human Factors, 25(1), 1-16. 1985
- 8 HART S G
HAUSER J R
LESTER P Inflight evaluation of four measures of pilot workload. Proceedings of the 28th Annual Meeting of the Human Factors Society. Santa Monica, CA: Human Factors Society, 732-736. 1984
- 9 ROSCOE A H Heart rate as an in-flight measure of pilot workload. Proceedings of the Workshop on Flight Testing to Identify Pilot Workload and Pilot Dynamics. (AFFTC-TR-82-5) Edwards AFB, CA, 338-349. 1982
- 10 HART S G
et al (in press) Type A versus Type B: Comparison of workload, performance, and cardiovascular measures.
- 11 WIERWILLE W W
RAHIMI M
CASALI J G Evaluation of 16 measures of mental workload using a simulated flight task emphasizing mediational activities. Human Factors, 27(5), 489-502. 1985
- 12 MULDER G Sinus arrhythmia and mental workload. In Mental Workload: Its Theory and Measurement. N Moray (Ed.) New York: Plenum Press, 327-344. 1979
- 13 ALLEN E M
DICEPOLA M N Mental workload in rule-based problem solving. Unpublished Masters Thesis. Toronto: University of Toronto. 1985
- 14 CHAN G
KRUSHELNYCKY E Mental workload in knowledge-based problem solving. Unpublished Masters Thesis. Toronto: University of Toronto. 1985
- 15 HART S G
STAVELAND L E (in press) Development of a multi-dimensional workload rating scale: Results of empirical and theoretical research. In P A Hancock & N Meshkati (Eds.) Human Mental Workload. Amsterdam: Elsevier.
- 16 HAWORTH L A
BIVENS C C
SHIVELY R J (in press) An investigation of single-piloted advanced cockpit and control configurations for nap-of-the-earth helicopter combat mission tasks. Proceedings of the 1986 Meeting of the American Helicopter Society, Washington, D C.



CHAPTER 16

INVESTIGATION OF WORKLOAD MEASURING TECHNIQUES:
A THEORETICAL AND PRACTICAL FRAMEWORK

by

René C van de Graaff
National Aerospace Laboratory NLR
Amsterdam
The Netherlands

ABBREVIATIONS

AGL	Above Ground Level
AP	Autopilot
ATC	Air Traffic Control
DME	Distance Measurement Equipment
FD	Flight Director
FT	Feet
ILS	Instrument Landing System
KT	Knot(s)
NDB	Non Directional Beacon
SYNC	Synchronizer
T/L	Take-off and Landing
VHF	Very High Frequency
VOR	VHF Omnidirectional Range
VTHR	Threshold Speed.

During the last two decades considerable research efforts have been devoted to developing a proper framework from which pilot workload can be analysed in a systematic manner. In this context, several attempts have been made to come to an unequivocal definition of the concept "workload" as well as to find an adequate method for its "measurement". At present it must be stated, however, that these efforts have not yet produced a satisfactory result. Although there is general agreement among investigators about the existence of different measures which could be used in some way as workload indicators, there is still no agreement about which these are, about which are the most effective, or about how (combinations of) these measures can be related to changes in the workload of an operator.

This situation has resulted in the development of a large number of measures for "quantifying" the operator's workload, whereas relatively little systematic research has been devoted to basic aspects, such as estimating the sensitivities of specific measures with respect to different task conditions and the relative practical usefulness of different techniques, within different operational environments. Furthermore, it has been assumed in a growing number of workload investigations that we are dealing with a multidimensional concept, so that a combination of measures is needed, or alternatively one measure with sensitivity to several dimensions, in order to come to a satisfactory evaluation of the operator's workload. This idea has had little impact, however, on the development and use of workload measurement techniques.

The foregoing outlines some reasons for developing a new approach towards the study of workload measurement. Such an approach should be based upon the presumption that the concept of workload encompasses various task- and operator-related aspects, for which each measure will most likely have a different sensitivity. In addition, the data obtained with these different measures must be integrated in a proper way in order to arrive at valid conclusions.

The following section discusses the implications of such an approach in detail. Subsequently, an experimental program is described, which could be used as a framework for systematic research on workload measurement techniques according to the notions mentioned above. This approach does not deal specifically with the development of new, independent measures but rather with the problem of how already existing measures can be used most effectively in a complex operational environment and how the results from several measures can be integrated to arrive at generally acceptable conclusions.

It should be noted that no attempt has been made here to propose the "correct" definition of the term "workload". Instead, it has been subsumed that this term pertains to a certain concept which cannot be decomposed satisfactorily in terms of its apparent components, but, when used as an operational concept, needs no further explicit explanation.

BASIC IDEAS AND AIMS

The basic ideas and aims beyond such a global research program can, as suggested above, be formulated more specifically as follows:

- 1 It is assumed that workload generally encompasses *several components*, such as time-stress, effort, etc. It is not clear which components play a part in a specific situation, nor what the impact of each is upon the overall perception of workload. It is therefore advocated that attention be paid to the sensitivity of specific workload measures to different aspects of the task.
- 2 As a further consequence of 1 it can be stated that research on workload measures should not be based upon any underlying assumptions about the existence of a superior method which can be used as a criterion (eg, "task complexity") for the

evaluation of other methods. Any such *a-priori* selection of task conditions according to a particular criterion, which may emphasize specific properties of a given task situation, could conceivably suppress the usefulness of certain other measures, due to an insufficient degree of variation of (other) task variables to which these measures are specifically sensitive. In other words, any assumption which supposes that "degrees" of workload can be indicated on the basis of one criterion exclusively is contradictory to the idea that workload is multidimensional. Investigations of workload measures, such as that proposed in the following, should therefore *not start from an a-priori ranking of task conditions* with respect to the expected "amount" of workload involved in the tasks. Instead, the underlying rationale for drawing certain conclusions should first carefully be considered.

3 Most workload studies focus exclusively on identifying *differences in workload*. It is important, however, that such studies focus also on the problem of identifying *similar workload levels* with respect to different tasks. This can have considerable importance for the experimental design. Equivalence in workload level has, for example, still not been demonstrated in cases where measured differences do not reach a significant level. It is necessary in workload studies, therefore, to establish the power of the intended statistical tests in an early stage of the investigation. Only in cases of sufficient power of the statistical tests being used (larger than, say, 0.7) is it possible to identify with a reasonable certainty both possible similarities and differences in workload for different task situations.

4 As a consequence of 3, *minimal differences of interest* ("difference margins") have to be specified with respect to each workload measure. The smaller these minimal differences are, the larger the sample-size must be in order to obtain the same power of the test. It is usually of little interest, for example, to detect "small" differences between two tasks, such as differences in average heart rate of, say, 0.1 beat per minute, or differences on a 10-point rating (interval-) scale between values of, say, 6.2 and 6.3. By thoughtfully selecting plausible difference margins, it should therefore be possible to use workload measures as indicators of differences as well as similarities in workload at an *a-priori* specified power. Clearly, there is a need for a universal agreement in connection with the specification of the indifference margins for workload measures*.

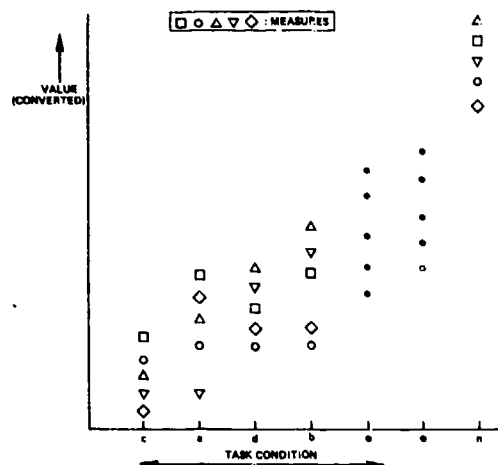


Fig. 1 Illustration of hypothetical experimental results for different modality measures (e.g. mean heart rate, subjective rating, etc.)

5 Since the data from a specific measure have indicated that either a (positive or negative) change or an equivalence of workload has occurred, it is necessary to define a *strategy* for drawing conclusions on the basis of a number of different modality measures, including some possibly contradictory results. A hypothetical situation is indicated schematically in Figure 1. The figure shows that:

- (i) Task conditions c and n are discriminated from each other by all measures in an unequivocal way.
- (ii) Task conditions d and b can not be discriminated from each other unequivocally by all the measures; that is, 3 of the 5 measures indicate a "positive" trend, while the remaining 2 indicate equivalence. The same observation holds for conditions c and a, where 4 of the 5 data indicate differences.
- (iii) The relative workload levels for the task conditions a and d, as indicated by the difference measures, are very contradictory.

It is obvious that case (iii) has to be examined more thoroughly. (Assuming that all measures involved are demonstrably valid workload indicators, one explanation is that artefacts have occurred in the data). Situation (i) does not evoke any interpretation problem; for comparable task situations it would seem sufficient to use one or two of the most convenient measures.

*In this context it is proposed here that it should, in first instance, be sufficient to accept an "indifference margin" of approximately 2.5 beats per minute for the average heart rate and of circa 0.3 point for a 10-point subjective rating scale (on an interval level).

For situation (ii), on the other hand, it should be concluded by *convention* whether or not a difference in workload has been observed. That is to say, when using several measures simultaneously, it should be acceptable to tolerate a small percentage of contradictory results (eg 20 percent) when formulating the final conclusion (alternatively, all deviation outcomes shall be inspected for the presence of artefacts, which could raise the cost of an investigation appreciably). In such cases the use of a proper weighting function with respect to the separate outcomes should also be considered. If a critical number of deviating results is exceeded, it is then necessary to carry out additional investigations of the considered cases until a desired level of homogeneity in the results has been obtained.

The problem of how to deal with contradictory results obtained from different measures is closely connected with the operational situation. That is, due to the complexity of most task situations it can usually not be postulated in advance which criterion should be used to draw final conclusions about the workload involved. Consequently, an operationally-based research program which includes various workload measures is advocated, on the basis of which a convention for drawing conclusions can be agreed upon. The ultimate objective of such a convention is to come to a generally acceptable framework for the evaluation of the outcomes arising from a set of separate measures.

6 The task situations to be selected for workload experiments must correspond to the ultimate complex *operational environment* for which the methods developed are intended to be used. This objective also supports the need, mentioned in 5 for an in-flight research program to be used as a common basis for different investigators for comparing and collating measures, evaluating strategies and effects of task aspects, etc. Such a program could be extended by including progressively more relevant task situations, ultimately arriving at a general framework from which operationally oriented workload studies can systematically proceed.

EXPERIMENTAL PROGRAM

The purpose of this section is to propose a basis for an in-flight experimental program which is in accordance with the ideas set forth in the foregoing. To accomplish this the experimental task conditions should be selected on the one hand on the basis of the irrelevancy with respect to operational situations and on the other hand also with respect to a-priori expectations about whether these conditions will produce a reasonable spread in the workload data. The analyses of the data obtained for the different measures will be aimed primarily at producing results which can be organised and presented in a form similar to that of Figure 1, from which one can proceed to the problem of drawing specific conclusions on the basis of different measures. Thereby, by comparing any systematic variations occurring in the different measures with the random variations occurring in such analyses, also an impression can be formed of the specific sensitivities of different measures.

The workload measures considered in connection with this experimental program must be selected on the basis of their expected "practical usefulness". By this is meant on one hand that the measures to be considered are expected to have a sufficient "discriminative power" with respect to the task conditions for which they are used, and on the other hand that these measures are expected to interfere as little as possible with the actual task. (Such a selection process can also be useful for clarifying specific pros and cons of certain measures with respect to operational applicability).

The experimental program set forth in the following deals with flight conditions for a fixed wing transport aircraft. The explanations are adopted from the findings of an in-flight study on pilot workload at the National Aerospace Laboratory (NLR), carried out in 1985-1986*. In addition to a discussion of some of the practical implications of such a program, an overview is given of some workload measuring techniques which are proposed as initial candidates for simultaneous investigation.

Task conditions

The heart of the program is the definition of the experimental task conditions. The problem faced here is to define a set of relevant tasks which cover a broad variety of operational flight conditions in a balanced way, but which do not exceed certain practical restrictions, such as time limits (ie, cost, pilot fatigue**), air traffic control restrictions, unacceptable risks, etc. Furthermore, as mentioned above, it is desired that the selected tasks give rise to a reasonable spread in the workload data to be obtained.

Taking these considerations into account, the following matrix of flight conditions for the civil aviation operating environment is proposed. The experimental tasks consist of flying procedural approaches (with an external view occluding visor), with each experimental run starting on "downwind", approximately 10 minutes before touchdown. The independent task variables are based on (i) the different approach aids, that is, ILS+FD, ILS, VOR+DME, NDB, (ii) the manner of pilot control, that is, automatic, manual, or manual with simulated trim malfunction (ie, retrimming prohibited after downwind), and (iii) the number of crew members, this is, 2 man versus 1 man crew. Because of time constraints it is advisable to carry out not more than 8 approaches within each experimental session. A proposed experimental design is presented in Table 1.

The different kinds of approaches are illustrated in Figures 2-4. The approaches consist of four compatible segments: downwind, turn/intercept, 1st segment final (above 1000 feet), 2nd segment final (below 1000 feet), which are very suitable for making comparisons between scenarios on the basis of the workload data obtained.

It is advisable to terminate each approach with an overshoot (at approximately 50 feet) as this will increase the flexibility of the program considerably, in addition to keeping the costs down. (Note that some of the approaches might conceivably have to be flown in opposite direction to the rest of the landing traffic due to local circumstances of wind and beacon locations).

* At the moment of submission of this paper to AGARD, the study is in progress.

** For obvious reasons, experiments are to be carried out with one subject, pilot (left seat) and one safety pilot.

TABLE 1
Experimental task conditions

Task Condition	Control Mode	Number of Crew Members	Approach Aid
a	Autopilot	2	ILS+FD/AP
b	Manual	2	ILS+FD
c	Manual	2	ILS
d	Manual	1	ILS
e	Manual	2	VOR+DME
f	Manual	2	NDB
g	Manual	1	NDB
h	Manual with trim malfunction	2	NDB

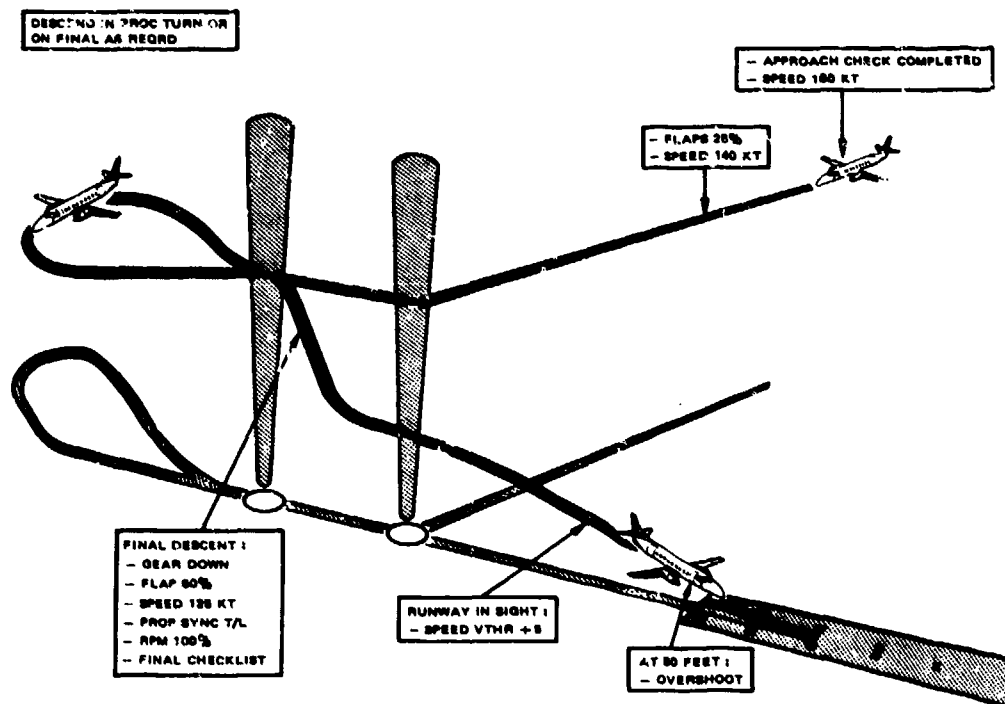


Fig. 2 NDB approach

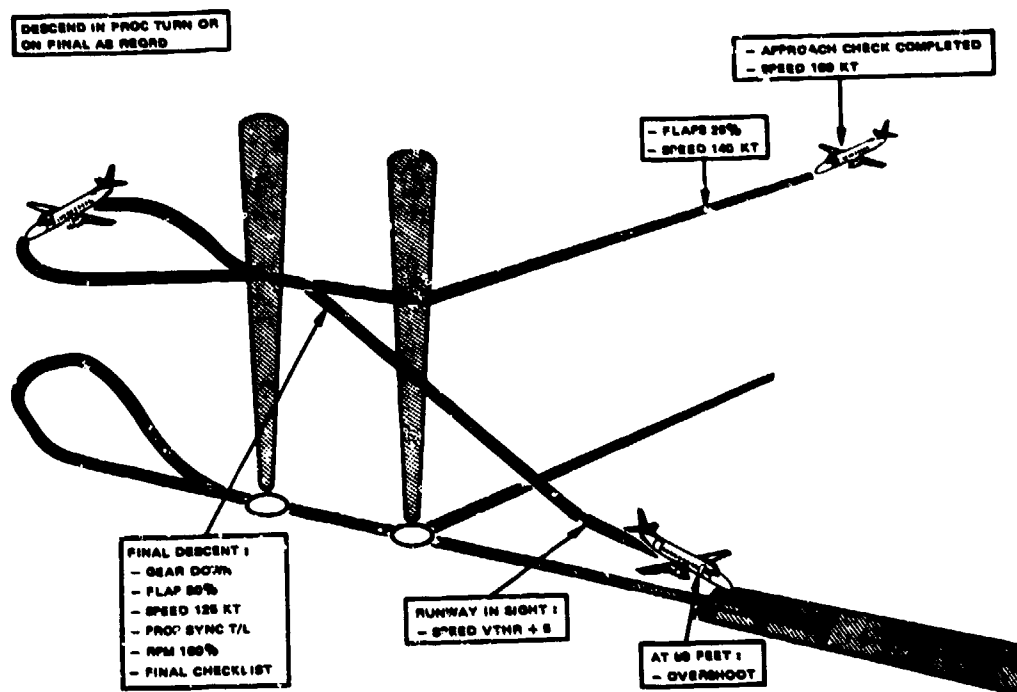


Fig. 3 VOR approach

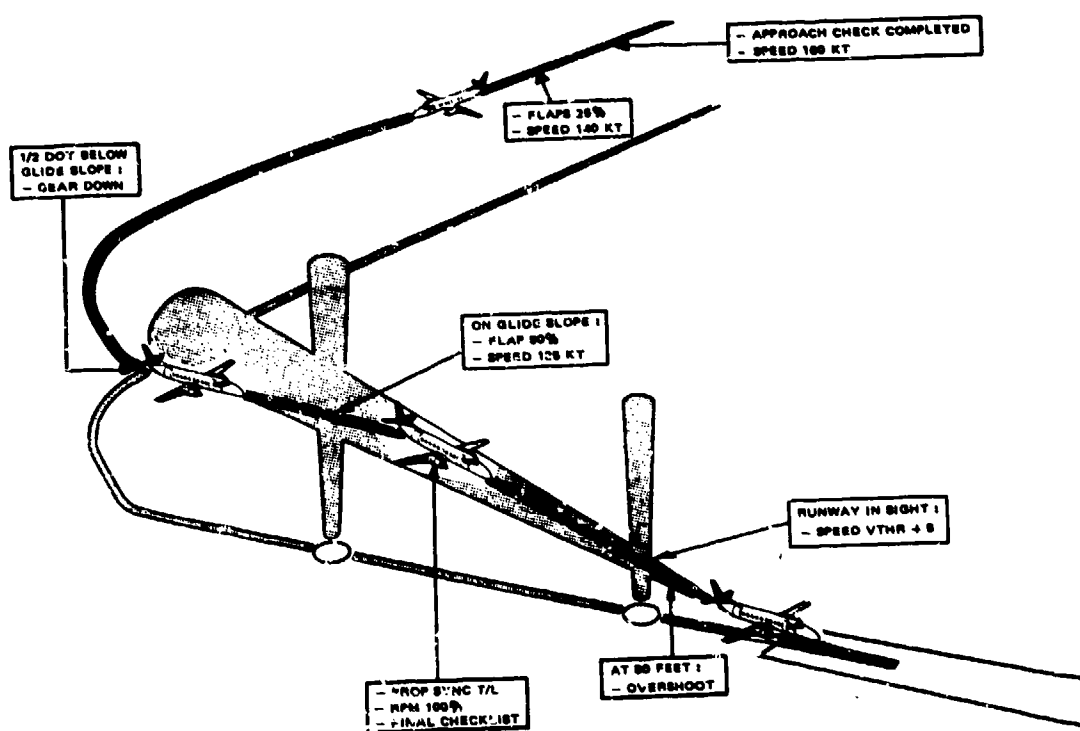


Fig. 4 ILS approach

Rating procedure

The suitable moment for subjective ratings (if ratings are to be given) is at the end of each segment, resulting in 4 ratings per approach. The ratings should concern the previously flown segment exclusively. Ratings can be obtained from both the subject pilot and the safety pilot. A proposed safe rating procedure (especially if the giving of ratings requires an appreciable amount of time) is as follows:

The safety pilot determines the appropriate moment, and says (after having given his own ratings): "I have control".

The subject pilot confirms with "You have control" and then gives his ratings, while the safety pilot flies the aircraft.

After completing the ratings, the subject-pilot takes over control, saying: "I have control" (Safety pilot confirms: "You have control").

The fourth rating (for the second segment of final) can be given during the overshoot. There is plenty of time during this flight segment, so that additional comments with respect to the entire foregoing approach can also be given.

Instructed performance standard

The following formulation of the requested performance standard is recommended as being relevant for real flight conditions: "You are requested to fly the procedures as accurately as possible, however, without violating your own standards (no exaggerations)".

Crew coordination

During the 2-man crew task conditions the safety pilot will perform the duties of the first officer; ie, he will take care of ATC communication, select beacons, set flaps on request, read out checklists and perform such other activities as would be expected from the first officer. In the 1-man crew task conditions, the subject-pilot must take over these duties in addition to the normal duties required for flying the aircraft.

Workload Measures

Given that the experimental set-up is designed to allow the simultaneous investigation of various measuring techniques (assuming that the experiment will be carried out on a well instrumented research aircraft), a careful selection has to be made of which specific techniques are to be considered within each experiment. Important for this selection will be, of course, the degree to which these measures interfere with the main flying task and with the other measurements. In Table 2 a number of proposed measures are listed (most of which are currently being considered in the above mentioned research program at the NLR). Some specific details about the measures in Table 2 are summarised as follows:

TABLE 2

Proposed workload measures to be considered within the study program

Pilot Ratings

1. McDonnell's 10-point demand scale (Ref. 1)
2. SWAT 3x3 rating matrix (Ref. 2)
3. Pre- and post-experimental ranking of task conditions

Heart Rate

1. Basic statistics (mean, standard deviation, root mean square, root mean square of successive differences)
2. Measures based on the spectral content of the signal

Primary Task Measures

1. Control activity
2. Task performance
3. Error frequencies

Model Measures

1. Control effort (Ref. 3)
2. Decision load (Ref. 4)

Other Measures

1. Time-motion parameters
2. Secondary task performance
3. Retrospective measures based on video replay

Two subjective rating techniques (McDonnell/SWAT) have been selected on the basis of their ability to produce ratings on an interval scale, in addition to the expectation, based on previous experience, that these two techniques can be used simultaneously without mutual interference. The McDonnell technique (1) involves pilot ratings, based upon the attentional demands of the task on a 10-point scale. The SWAT technique (2) involve separate ratings with respect to time, effort and psychological stress aspects of the task, in order to arrive at one single assessment of the pilot's overall workload. In addition to the ratings given during the experiment, the subjects are requested to rank the different approaches according to the expected/experienced task workload also before and after the experiment.

In choosing heart rate as a workload measure it is worthwhile to note the advantage of its objectivity, the ease of recording it and the non-intrusive nature of its measurement.

The primary task measures (for each flight segment) involve a number of relevant statistics related to the pilot's control activity (for the manually flown tasks) and flying performance.

The data obtained make it possible also to investigate the usefulness of predictions of pilot workload based upon mathematical models of pilot-aircraft interactions. Especially, for the manually flown tasks the so-called "control effort" parameter (E) mentioned in reference (3) seems worthwhile to be investigated further. This parameter indicates, among others, the sensitivity of task performance to a model parameter reflecting level-of-attention. Also model-based parameters reflecting "busyness" aspects or "decision-loading" aspects, such as the "Expected Net Gain for Procedure execution (ENGP), mentioned in reference (4) can also be investigated further. It should be noted that such use of modelling presupposes the availability of an adequate system/aircraft model.

Finally, crew activity is to be recorded on video in order to enable various task analyses, investigations of pilot errors, and other further analyses based on video replay.

Final Remarks

As is known from other in-flight experiments, a relatively large variability in data can be expected due to such uncontrollable factors as atmospheric conditions and airport traffic. Therefore it is advisable to fly all approaches at one specific airport. In the current NLR research program, a preliminary study has indicated that 20 sessions (including 20x8=160D approaches) are necessary to obtain an adequate level of statistical reliability in the experiment.

SUMMARY AND CONCLUSIONS

Workload research has lead in the past to the development of various measures, mostly concerning different aspects of task workload, in a separate and isolated way. In addition, present opinion assumes more and more that, in order to achieve a satisfactory workload evaluation, a matrix of measures is needed.

This paper discusses a number of considerations involved in the setting up of an investigation dealing with the problem of being able to draw conclusions from a variety of experimental measures in a complex task situation. Several implications are pointed out, such as the problem of dealing with contradictory outcomes, the designating of artefacts, and the problem of formulating final conclusions without the (a-priori) availability of a superior method for evaluating other methods. Finally, an experimental program is outlined which is based on (normal) approach conditions for civil fixed wing aircraft. The task conditions in this experiment are selected to serve as an operationally based framework for comparing different workload evaluation methods, for evaluating the effects of specific task conditions and for investigating the strategies needed for drawing final conclusions from a variety of outcomes.

REFERENCES

- 1 MCDONNELL J D Pilot rating techniques for the estimation and evaluation of handling qualities. AFFDL-TR-68-76, 1968
- 2 REID G B
SHINGLEDECKER C A
NYGREN T E
EGGEMEIER F T Development of multidimensional subjective measures of workload. Proceedings of the 1981 IEEE International Conference on Cybernetics and Society, pp 403-406, 1981
- 3 van de GRAAFF R C NLR research on pilot dynamics and workload. Proceedings of the Workshop in Flight Testing to Identify Pilot Workload and Pilot Dynamics, AFFTC-TR-82-5, pp 79-90, 1982
- 4 MILGRAM P
van de WINGAART R F
VEERBEEK H
BLEEKER O F Multi-crew model analytic assessment of decision-making demand and landing performance. Twentieth Annual Conference on Manual Control. NADA-CP-2341, Volume II, pp 373-396, 1984

APPENDIX 1

**FINAL FIVE MINUTES OF A MANUALLY FLOWN ILS APPROACH AND LANDING OF A
TWO-PILOT PASSENGER JET TRANSPORT (USING FLIGHT DIRECTOR)
WEATHER: 200 FT CLOUD BASE. RAIN, RVR 700m**

Approx Time to TD	Distance From TD	Speed KIAS	Activity
4½	14	250	Descending through 4,000 ft in thick icing cloud to level at 2,000 ft. (Descent checks completed by TOD.) Speedbrake out. Autopilot disengaged. Radio nav aids already set and identified for landing
		210	
	13	210	Through 3,000 ft. Radar vectors to ILS localiser. Approach checklist.
	11	210	Speedbrake cancelled.
			Level off at 2,000 ft. Flaps to 1. Radio call to 'approach control'.
		190	Altimeters — P1 sets QFE. P2 set QNH.
			Flaps to 5
		170	On course to intercept localiser.
3	8	170	Localiser intercepted — Radio change to tower frequency. Localiser established.
	6	170	Landing gear down. Flaps to 15.
		150	Landing checks
2½	5	140	Glide slope capture
		VAT+10	Land flap selected checks complete.
			Descending on glide slope.
2	4	VAT+10	Outer Marker — Height check. Land clearance and surface W/V and RVR passed by control.
			500 ft QFE Height calls and incapacitation check. Speed and rate of descent monitored.
			300 ft QFE 100 ft above call.
			200 ft Decision height. 'Approach lights' call.
0			Flare and touchdown.
			Reverse thrust — deceleration
+½		60K	Reverse thrust cancelled — braking gently.
			Nose-wheel steering.

NB Performance limits to be clearly defined eg 3K IAS, 50 ft/min rate of descent on G/S. Touchdown within 100m of 'numbers' at < 6 ft/sec.

P1 — Pilot flying

P2 — Co-pilot

APPENDIX 2

FIVE MINUTE RECCE/ATTACK TASK FOR FAST JET AIRCRAFT (SINGLE PILOT)

1. CHOICE OF TARGET FOR RECCE/ATTACK

A realistic target would be a simple, soft skinned, vehicle parked on a two lane track, at known position, in flat/undulating open country.

2. ROUTE AND PARAMETERS FOR ATTACK TASK

Route comprises a single 3 minute navigation leg to IP followed by 1 min IP to target run, positioning for standard 45 degree tip-in a shallow 5-8 degree dive. Parameters of tip-in according to aircraft type, following attack, defensive recovery manoeuvre and fix before withdrawal, 1 minute after target. (The recce task would involve a similar sequence but camera selection would replace weapon selection and a level off set fly-by would replace the tip-in attack.)

3. ACCURACY

EN ROUTE AND IP RUN $\pm 20 \text{ kt} \pm 100 \text{ ft}$ at peacetime minimum height or $\pm 200 \text{ ft}$ above 500 ft AGL.

Define pull-up position relative to target

Pull-up point $\pm 300 \text{ m}$ laterally, $\pm 3 \text{ sec}$ along track

Define dive angle $\pm 1 \text{ degree}$

3 sec tracking to release point

Release at $\pm 20 \text{ kt}$ planned speed

Pull-off target 4g. Defensive/position manoeuvre to roll out towards next turning point.

4. DETAILS OF ATTACK TASK

a. Approach to IP

Whilst flying within set parameters, complete following actions:

Check slip ball and trim out sideslip at planned attack speed

Weapon switching up to final arming switch

1 x track-check, map-to-ground

1 x revision ETA for IP $\pm 5 \text{ sec}$ (10)?

8 x simulated checks of wing-man's six o'clock high

Estimate (or check wind from INS) W/S for weapon release

Set WIND/AM Depression for Main or reversionary attack

Carry out height fix to IP to update Pressure alt/or auto height fix at IP(INAS)

Set or confirm next heading

Acquire IP visually

b. IP to Pull-Up

Whilst flying within set parameters:

Complete height fix in INAS equipped aircraft

Update INAS

Accelerate to Attack Speed

Check track to pull-up point and pull-up time

Adjust Sighting (Call up attack picture on HUD).

c. Pull-up to Weapon Release

Initiate Pull-Up

Acquire target

Top at height required to achieve planned dive angle

Check speed/power

Sight or Bomb Fall Line on Target (3 sec)

Make Final arming switch

Start Camera (if not automatic)
Phase Change (if INAS equipped)
Track target for 3 sec up to weapon release
Release Weapon.

d. **Escape**

Recover from dive
Make defensive manoeuvre
Put weapon switches safe
Track to next turning point
Locate and identify other aircraft (No 2) if available
Switch off camera
Regain HUD NAV mode (if not automatic).

APPENDIX 3

BROAD OUTLINE OF A STANDARD HELICOPTER TASK

Total time — 5 minutes

Scenario reasonably well detailed eg 'nap of earth' flying for recce task.

0	take-off — spot turn
0.30	transition — climb to 1250 feet
2.00	cruise for one minute
3.30	descent to nap of earth for precision low level observation task at high speed covering a 'figure of 8' pattern.
4.30	approach to land
5.00	land

NB Further details of aircraft parameters and of scenario to be added as necessary.

REPORT DOCUMENTATION PAGE			
1. Recipient's Reference	2. Originator's Reference	3. Further Reference	4. Security Classification of Document
	AGARD-AG-282	ISBN 92-835-1546-3	UNCLASSIFIED
5. Originator	Advisory Group for Aerospace Research and Development North Atlantic Treaty Organization 7 rue Ancelle, 92200 Neuilly sur Seine, France		
6. Title	THE PRACTICAL ASSESSMENT OF PILOT WORKLOAD		
7. Presented at			
8. Author(s)/Editor(s)	Editor: Alan H. Roscoe, MD		9. Date June 1987
10. Author's/Editor's Address	Britannia Airways Limited, Luton Airport Bedfordshire LU2 9ND, UK		11. Pages 140
12. Distribution Statement	This document is distributed in accordance with AGARD policies and regulations, which are outlined on the Outside Back Covers of all AGARD publications.		
13. Keywords/Descriptors	<p>Work measurement Pilots (personnel) Performance evaluation</p>		
14. Abstract	<p>Whether one is attempting to reduce workload in the cockpit of a combat aircraft to improve mission effectiveness, or to optimise workload levels on the flight deck of a civil airliner to improve safety, it is important to be able to assess workload in practical terms. In the case of the civil transport aircraft the findings of the President's Task Force on Crew Complement have underlined the need to assess workload in flight reliably in order to satisfy certification requirements for new aircraft.</p> <p>The main purpose of this multi-author AGARDograph is to provide guidance for the reader who may wish to assess pilot workload in practical situations. It is not meant to be a comprehensive treatise on the subject.</p> <p>Chapter 1 introduces the subject and reviews briefly the various techniques available for assessing pilot workload. Seven of the remaining fifteen chapters describe techniques that have been used successfully in flight; four chapters are concerned primarily with assessing workload for the purpose of aircraft certification; and other chapters discuss techniques that show promise and with further development may well become available for practical use.</p> <p>This AGARDograph has been prepared at the request of the Flight Mechanics Panel of AGARD.</p>		

<p>AGARDograph No.282 Advisory Group for Aerospace Research and Development, NATO THE PRACTICAL ASSESSMENT OF PILOT WORK-LOAD Edited by Alan H. Roscoe, MD Published June 1987 140 pages</p> <p>Whether one is attempting to reduce workload in the cockpit of a combat aircraft to improve mission effectiveness, or to optimise workload levels on the flight deck of a civil airliner to improve safety, it is important to be able to assess workload in practical terms. In the case of the civil transport aircraft the findings of the President's Task Force on Crew Complement have underlined the need to</p> <p>P.T.O.</p>	<p>AGARD-AG-282</p> <p>Work measurement Pilots (personnel) Performance evaluation</p>	<p>AGARDograph No.282 Advisory Group for Aerospace Research and Development, NATO THE PRACTICAL ASSESSMENT OF PILOT WORK-LOAD Edited by Alan H. Roscoe, MD Published June 1987 140 pages</p> <p>Whether one is attempting to reduce workload in the cockpit of a combat aircraft to improve mission effectiveness, or to optimise workload levels on the flight deck of a civil airliner to improve safety, it is important to be able to assess workload in practical terms. In the case of the civil transport aircraft the findings of the President's Task Force on Crew Complement have underlined the need to</p> <p>P.T.O.</p>	<p>AGARD-AG-282</p> <p>Work measurement Pilots (personnel) Performance evaluation</p>
<p>AGARDograph No.282 Advisory Group for Aerospace Research and Development, NATO THE PRACTICAL ASSESSMENT OF PILOT WORK-LOAD Edited by Alan H. Roscoe, MD Published June 1987 140 pages</p> <p>Whether one is attempting to reduce workload in the cockpit of a combat aircraft to improve mission effectiveness, or to optimise workload levels on the flight deck of a civil airliner to improve safety, it is important to be able to assess workload in practical terms. In the case of the civil transport aircraft the findings of the President's Task Force on Crew Complement have underlined the need to</p> <p>P.T.O.</p>	<p>AGARD-AG-282</p> <p>Work measurement Pilots (personnel) Performance evaluation</p>	<p>AGARDograph No.282 Advisory Group for Aerospace Research and Development, NATO THE PRACTICAL ASSESSMENT OF PILOT WORK-LOAD Edited by Alan H. Roscoe, MD Published June 1987 140 pages</p> <p>Whether one is attempting to reduce workload in the cockpit of a combat aircraft to improve mission effectiveness, or to optimise workload levels on the flight deck of a civil airliner to improve safety, it is important to be able to assess workload in practical terms. In the case of the civil transport aircraft the findings of the President's Task Force on Crew Complement have underlined the need to</p> <p>P.T.O.</p>	<p>AGARD-AG-282</p> <p>Work measurement Pilots (personnel) Performance evaluation</p>

<p>assess workload in flight reliably in order to satisfy certification requirements for new aircraft.</p> <p>The main purpose of this multi-author AGARDograph is to provide guidance for the reader who may wish to assess pilot workload in practical situations. It is not meant to be a comprehensive treatise on the subject.</p> <p>Chapter 1 introduces the subject and reviews briefly the various techniques available for assessing pilot workload. Seven of the remaining fifteen chapters describe techniques that have been used successfully in flight; four chapters are concerned primarily with assessing workload for the purpose of aircraft certification; and other chapters discuss techniques that show promise and with further development may well become available for practical use.</p> <p>This AGARDograph has been prepared at the request of the Flight Mechanics Panel of AGARD.</p> <p>ISBN 92-835-1546-3</p>	<p>assess workload in flight reliably in order to satisfy certification requirements for new aircraft.</p> <p>The main purpose of this multi-author AGARDograph is to provide guidance for the reader who may wish to assess pilot workload in practical situations. It is not meant to be a comprehensive treatise on the subject.</p> <p>Chapter 1 introduces the subject and reviews briefly the various techniques available for assessing pilot workload. Seven of the remaining fifteen chapters describe techniques that have been used successfully in flight; four chapters are concerned primarily with assessing workload for the purpose of aircraft certification; and other chapters discuss techniques that show promise and with further development may well become available for practical use.</p> <p>This AGARDograph has been prepared at the request of the Flight Mechanics Panel of AGARD.</p> <p>ISBN 92-835-1546-3</p>
<p>assess workload in flight reliably in order to satisfy certification requirements for new aircraft.</p> <p>The main purpose of this multi-author AGARDograph is to provide guidance for the reader who may wish to assess pilot workload in practical situations. It is not meant to be a comprehensive treatise on the subject.</p> <p>Chapter 1 introduces the subject and reviews briefly the various techniques available for assessing pilot workload. Seven of the remaining fifteen chapters describe techniques that have been used successfully in flight; four chapters are concerned primarily with assessing workload for the purpose of aircraft certification; and other chapters discuss techniques that show promise and with further development may well become available for practical use.</p> <p>This AGARDograph has been prepared at the request of the Flight Mechanics Panel of AGARD.</p> <p>ISBN 92-835-1546-3</p>	<p>assess workload in flight reliably in order to satisfy certification requirements for new aircraft.</p> <p>The main purpose of this multi-author AGARDograph is to provide guidance for the reader who may wish to assess pilot workload in practical situations. It is not meant to be a comprehensive treatise on the subject.</p> <p>Chapter 1 introduces the subject and reviews briefly the various techniques available for assessing pilot workload. Seven of the remaining fifteen chapters describe techniques that have been used successfully in flight; four chapters are concerned primarily with assessing workload for the purpose of aircraft certification; and other chapters discuss techniques that show promise and with further development may well become available for practical use.</p> <p>This AGARDograph has been prepared at the request of the Flight Mechanics Panel of AGARD.</p> <p>ISBN 92-835-1546-3</p>

AGARD

NATO OTAN

7 rue Ancelle - 92230 NEUILLY-SUR-SEINE
FRANCE

Telephone (1) 47.38.67.30 - Telex 610 178

**DISTRIBUTION OF UNCLASSIFIED
AGARD PUBLICATIONS**

AGARD does NOT hold stocks of AGARD publications at the above address for general distribution. Initial distribution of AGARD publications is made to AGARD Member Nations through the following National Distribution Centres. Further copies are sometimes available from these Centres, but if not may be purchased in Microfiche or Photocopy form from the Purchase Agencies listed below.

NATIONAL DISTRIBUTION CENTRES

BELGIUM

Coordonnateur AGARD -- VSL
E: Major de la Force Aérienne
C: Rue Reine Elisabeth
B: d'Evere, 1140 Bruxelles

CANADA

Defence Scientific Information Services
Dept of National Defence
Ottawa, Ontario K1A 0K2

DENMARK

Danish Defence Research Board
Ved Idmatsparken 4
2100 Copenhagen Ø

FRANCE

O.N.E.R.A. (Direction)
29 Avenue de la Division Leclerc
92320 Châtillon

GERMANY

Fachinformationszentrum Energie,
Physik, Mathematik GmbH
Kernforschungszentrum
D-7514 Eggenstein-Leopoldshafen

GREECE

Hellenic Air Force General Staff
Research and Development Directorate
Hoflora, Athens

ICELAND

Director of Aviation
c/o Flugrad
Reykjavik

ITALY

Aeronautica Militare
Ufficio del Delegato Nazionale all'AGARD
3 Piazza Adenauer
00144 Roma/EUR

LUXEMBOURG

See Belgium

NETHERLANDS

Netherlands Delegation to AGARD
National Aerospace Laboratory, NLR
P.O. Box 126
2600 AC Delft

NORWAY

Norwegian Defence Research Establishment
Attn: Biblioteket
P.O. Box 25
N-2007 Kjeller

PORTUGAL

Portuguese National Coordinator to AGARD
Gabinete de Estudos e Programas
CLAPA
Base de Alfragide
Alfragide
2700 Amadora

TURKEY

Milli Savunma Bakanligi (MSB)
ARGE Daire Baskanligi (AROE)
Ankara

UNITED KINGDOM

Defence Research Information Centre
Kerrigern House
65 Brown Street
Glasgow G2 8EX

UNITED STATES

National Aeronautics and Space Administration (NASA)
Langley Research Center
M/S 180
Hampton, Virginia 23665

THE UNITED STATES NATIONAL DISTRIBUTION CENTRE (NASA) DOES NOT HOLD STOCKS OF AGARD PUBLICATIONS, AND APPLICATIONS FOR COPIES SHOULD BE MADE DIRECT TO THE NATIONAL TECHNICAL INFORMATION SERVICE (NTIS) AT THE ADDRESS BELOW.

PURCHASE AGENCIES

National Technical
Information Service (NTIS)
5285 Port Royal Road
Springfield
Virginia 22161, USA

ESA/Information Retrieval Service
European Space Agency
10, rue Mario Nikis
75015 Paris, France

The British Library
Document Supply Division
Boston Spa, Wetherby
West Yorkshire LS23 7BQ
England

Requests for microfiche or photocopies of AGARD documents should include the AGARD serial number, title, author or editor, and publication date. Requests to NTIS should include the NASA accession report number. Full bibliographical references and abstracts of publications are given in the following journals:

Scientific and Aerospace Information (SCI) (AR)
published by NASA Scientific and Technical
Information Branch
NASA Headquarters (NIT-40)
Washington D.C. 20546, USA

Government Report Announcements (GRA)
published by the National Technical
Information Service, Springfield
Virginia 22161, USA



Specialist Printing Services Limited
Jugwell Lane, Loughton, Essex IG10 3TZ

ISBN 92-875-1546-3